Exponential Convergence and stability of Howards's Policy Improvement Algorithm for Controlled Diffusions

EXPONENTIAL CONVERGENCE AND STABILITY OF HOWARD'S POLICY IMPROVEMENT ALGORITHM FOR CONTROLLED DIFFUSIONS

B. KERIMKULOV, D. ŠIŠKA, AND Ł. SZPRUCH

Abstract. Optimal control problems are inherently hard to solve as the optimization must be performed simultaneously with updating the underlying system. Starting from an initial guess, Howard's policy improvement algorithm separates the step of updating the trajectory of the dynamical system from the optimization and iterations of this should converge to the optimal control. In the discrete space-time setting this is often the case and even rates of convergence are known. In the continuous space-time setting of controlled diffusion the algorithm consists of solving a linear PDE followed by a maximization problem. This has been shown to converge; in some situations, however no global rate is known. The first main contribution of this paper is to establish global rate of convergence for the policy improvement algorithm and a variant, called here the gradient iteration algorithm. The second main contribution is the proof of stability of the algorithms under perturbations to both the accuracy of the linear PDE solution and the accuracy of the maximization step. The proof technique is new in this context as it uses the theory of backward stochastic differential equations.

1. Introduction

Stochastic control problems arise naturally in a range of applications in engineering, economics, and finance. Apart from very specific cases such as linear-quadratic control in engineering or the Merton portfolio optimization task in finance, stochastic control problems typically have no closed form solutions and have to be solved numerically. In this paper we consider the policy iteration algorithm and gradient iteration algorithm; see Algorithms 1 and 2. These are effectively a linearization method for the inherently nonlinear problem and play an essential role in numerical solutions of stochastic control problems.

We will consider the continuous space, continuous time problem where the controlled system is modeled by an $\mathbb{R}^d$-valued diffusion process. Let W be a d'-dimensional Wiener martingale on a filtered probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\geq 0}, \mathbb{P})$. Let us fix a finite time $T \in (0, \infty)$ and consider the controlled SDE

$$dX_s = b^\alpha(s, X_s)\, ds + \sigma(s, X_s)\, dW_s\,,\ s \in [t, T]\,,\ \ X_t = x\,. \tag{1}$$

Maxwell Institute Graduate School in Analysis and its Applications, Edinburgh, UK.

School of Mathematics, University of Edinburgh and Vega Protocol

School of Mathematics, University of Edinburgh and Alan Turing Institute

E-mail addresses: B.Kerimkulov@sms.ed.ac.uk, D.Siska@ed.ac.uk, L.Szpruch@ed.ac.uk.

Date: 25th May 2020,

2010 *Mathematics Subject Classification.* 93E20, 60H30, 65N12, 49L20.

Key words and phrases. Policy Improvement Algorithm, Stochastic Control, Backward Stochastic Differential Equation.

Supported by the Maxwell Institute Graduate School in Analysis and its Applications, a Centre for Doctoral Training funded by the UK Engineering and Physical Sciences Research Council (grant EP/L016508/01), the Scottish Funding Council, Heriot-Watt University and the University of Edinburgh.

Here $\alpha = (\alpha_s)$ is a control belonging to the space of admissible controls $\mathcal{A}$, valued in $A \subseteq \mathbb{R}^m$, and we will write $X^{t,x,\alpha}$ to denote the solution of (1) which starts from x at time t while being controlled by α. We shall consider the gain functional in the form

$$J(t,x,\alpha) := \mathbb{E}\left[\int_t^T f^\alpha(s, X_s^{t,x,\alpha})ds + g(X_T^{t,x,\alpha})\right] \tag{2}$$

for all $(t,x) \in [0,T] \times \mathbb{R}^d$ and $\alpha \in \mathcal{A}$. The value function $v = v(t,x)$ is given for all $t \in [0,T]$ and $x \in \mathbb{R}^d$ by

$$v(t,x) = \sup_{\alpha \in \mathcal{A}} J(t,x,\alpha). \tag{3}$$

We wish to solve the optimization problem, i.e., to find either the value function v or the optimal control α^* which achieves the maximum (or, if the supremum cannot be reached by $\alpha \in \mathcal{A}$, then an ε-optimal control $\alpha^\varepsilon \in \mathcal{A}$ such that $v(t,x) \leq J(t,x,\alpha^\varepsilon) + \varepsilon$). It is well known that (see, e.g., Krylov [6]) that under reasonable assumptions the value function satisfies the Bellman PDE:

$$\begin{aligned} \partial_t v + \frac{1}{2}\mathrm{tr}(\sigma\sigma^\top D_x^2 v) + \sup_{a \in A}(b^a D_x v + f^a) = 0 \text{ on } [0,T) \times \mathbb{R}^d, \\ v(T,x) = g(x) \ \text{ on } x \in \mathbb{R}^d. \end{aligned} \tag{4}$$

Moreover (again see Krylov [6]), it is sufficient to consider Markovian controls, i.e., processes $\alpha_s = a(s, X_s^{t,x,\alpha})$ for some measurable function $a : [0,T] \times \mathbb{R}^d \to A$. Thus if we have obtained the value function, then we can find the optimal control (if it exists) as

$$a^*(t,x) = \arg\max_{a \in A}\left(b^a(t,x)(D_x v)(t,x) + f^a(t,x)\right).$$

It is rarely possible to find a closed form solution to (4) and so various approximations have to be employed. One may, for example, choose to use a finite difference method to discretize (4) and indeed this has been widely studied; see, e.g., [12] or [14] and references therein. This results in a high dimensional nonlinear system of equations that still retains the structure of (4). To solve this nonlinear system one may apply the Howard's policy improvement algorithm. The rate of convergence would then follow from results available on discrete space-time control problems. However, to check that the assumptions required for convergence are satisfied is not straightforward and moreover it is dependent on the discretization scheme used.

An alternative approach is to linearize (4) and to iterate. The classical approach is the Bellman–Howard policy improvement/iteration algorithm. The algorithm is initialized with a "guess" of the Markovian control. Given a Markovian control strategy at step n one solves a linear PDE with the given control fixed and then one uses the solution to the linear PDE to update the Markovian control. In this paper we will show that this policy improvement algorithm (see Algorithm 1) and a variant which we call the gradient iteration algorithm (see Algorithm 2) converge, under appropriate assumptions, exponentially fast.

Iterative algorithms for the solution of optimal control problems go back to the work of Bellman [1, 2] where the value iteration algorithms for finite space-time problems are developed and their convergences are shown. Howard [3] proposed the policy improvement algorithm in the context of the discrete space-time Markovian decision rocess. Puterman and Brumelle's paper [4] was one of the first results on the convergence properties for the policy iteration for MDP problems. The abstract function space setting employed in the paper applies to both discrete and continuous settings. Their main observation is that the policy iteration can be viewed as a type of Newton's method. Hence similar convergence results to those known for Newton's method follow: in particular, if the initial guess is in a neighborhood of

the true solution, then the convergence will be quadratic. Puterman [5] applied this in a setting very similar to that of this paper to prove quadratic convergence in the neighborhood of the limit. Santos and Rust [9] consider the discrete time but continous space and controls setting. They extend the results of Puterman and Brumelle [4] to show global convergence, but without global rate, and quadratic local convergence rate of policy iteration and superlinear local convergence under more general conditions. In the case of stochastic control problems with jump-diffusion processes, Bäuerle and Rieder [17] have proved a convergence result of the Howard's policy improvement algorithm with the help of martingale techniques. In the fully discrete space and time setting Bokanowski, Maroso, and Zidani [13] have shown global superlinear convergence, under a monotonicity assumption on the matrices defining the control problem. Convergence of policy iteration has been recently proved by Jacka and Mijatović [18] and Jacka, Mijatović, and Siraj [19]. Further, Maeda and Jacka [20] have shown quadratic local convergence of the policy iteration algorithm for the time-independent control problem. The local quadratic convergence is similar to the result of Puterman [5] but the specific control problem is different and moreover they employ a completely different technique based on Schauder estimates for linear PDEs.

The main contributions of this paper are to establish a global rate of convergence and stability for the policy iteration algorithm and a variant, which we call the gradient iteration algorithm. The analysis is carried out using backward stochastic differential equations (BSDEs) and to the best knowledge of the authors this is the first time BSDEs have been used to study convergence of the policy iteration algorithm. The assumptions required for this are effectively Lipschitz dependence in the drift, diffusion, instantaneous payoff, and terminal payoff functions and independence of the diffusion matrix on the control; see (1). The stability results show that the policy iteration remains stable even if the linear PDE is solved only approximately and even if the maximization is step performed approximately. Moreover they allow one to devise computationally efficient algorithms as they show that in the initial steps it is sufficient to solve the linear PDE with very low accuracy, and a highly accurate PDE solver is only required for the final few iterations of the algorithms.

The paper is organized as follows. In Section 2 we introduce all the assumptions and notation used throughout the paper. In Sections 3 and 4 we state and prove the results concerning convergence of the gradient iteration algorithm and the policy improvement algorithm, respectively. Section 5 justifies the name "policy improvement algorithm" in that it shows that the value functions increase monotonically with iterations and it also shows that the algorithm converges under weaker assumptions than those required for obtaining the rate. Sections 6 and 7 prove the stability of the algorithms. In Section 8 we present an example that fits the setting of this paper. Finally, in Appendix A, we collect several known results from the theory of BSDEs that are essential for the proofs.

Algorithm 1 Policy improvement algorithm.

Initialization: make a guess of the control $a^0 = a^0(t,x)$.
while difference between v^{n+1} and v^n is large **do**
Given a control $a^n = a^n(t,x)$ solve the *linear* PDE

$$\begin{aligned} \partial_t v^n + \frac{1}{2}\mathrm{tr}(\sigma\sigma^\top D_x^2 v^n) + b^{a^n} D_x v^n + f^{a^n} = 0 \text{ on } [0,T)\times\mathbb{R}^d\,, \\ v^n(T,\cdot) = g \ \text{ on } x\in\mathbb{R}^d\,. \end{aligned} \tag{5}$$

Update the control

$$a^{n+1}(t,x) = \arg\max_{a\in A}\left[\left(b^a D_x v^n + f^a\right)(t,x)\right]\,. \tag{6}$$

end while
return v^n, a^{n+1}.

Algorithm 2 Gradient iteration algorithm.

Initialization: make a guess of the value function $v^0 = v^0(t,x)$.
while difference between v^n and v^{n-1} is large **do**
Given value function $v^{n-1} = v^{n-1}(t,x)$ update the control

$$a^n(t,x) = \arg\max_{a\in A}\left[\left(b^a D_x v^{n-1} + f^a\right)(t,x)\right]\,. \tag{7}$$

Solve the *linear* PDE

$$\begin{aligned} \partial_t v^n + \frac{1}{2}\mathrm{tr}(\sigma\sigma^\top D_x^2 v^n) + b^{a^n} D_x v^{n-1} + f^{a^n} = 0 \text{ on } [0,T)\times\mathbb{R}^d\,, \\ v^n(T,\cdot) = g \ \text{ on } x\in\mathbb{R}^d\,. \end{aligned} \tag{8}$$

end while
return v^n, a^n.

We would like to emphasize that Algorithm 1 and Algorithm 2 are different, although they look rather similar. In Algorithm 1, v^n is the value function for the Markov control a^n, since it solves the PDE (5). In Algorithm 2 v^n is not the value function for the Markov control a^n. This is due to the term $b^{a^n} D_x v^{n-1}$ in the linear PDE (8).

2. Assumptions and Notation

We fix a finite horizon $T \in (0,\infty)$. We assume that for some $m\in\mathbb{N}$ we have $A \subseteq R^m$ such that $0 \in A$. This is the space where the control processes α take values. We fix a filtered probability space $(\Omega, \mathcal{F}, \mathbb{F} = (\mathcal{F}_t)_{0\le t\le T}, \mathbb{P})$. Let $W = (W_t)_{t\in[0,T]}$ be a d'-dimensional Wiener martingale on this space. Moreover, we have the following:

(i) For $\gamma > 0$ and a predictable process ϕ let us define

$$\|\phi\|_{\mathbb{H}^2_\gamma} := \left(\mathbb{E}\int_0^T e^{\gamma s}|\phi_s|^2\,ds\right)^{\frac{1}{2}}\,.$$

For $\gamma = 0$ we will write $\|\cdot\|_{\mathbb{H}^2}$. We will use $\mathbb{H}^2$ to denote the set of all predictable processes ϕ such that $\|\phi\|_{\mathbb{H}^2} < \infty$. Note that the norm $\|\cdot\|_{\mathbb{H}^2}$ is equivalent to the norm $\|\cdot\|_{\mathbb{H}^2_\gamma}$ for any $\gamma \ge 0$.

(ii) Let $\mathcal{S}^2$ be the set of real valued $\mathbb{F}$-adapted continuous processes ϕ on $[0,T]$ such that
$$\|\phi\|_{\mathcal{S}^2} := \mathbb{E}\left[\sup_{0\le r\le T}|\phi_r|^2\right] < \infty\,.$$

(iii) For adapted processes ϕ such that $\int_0^t |\phi_s|^2\,ds < \infty$ almost surely we will define
$$(\phi \bullet W)_t := \int_0^t \phi_s\,dW_s\,.$$

(iv) For any continuous local martingale M let with $(\langle M\rangle_t)_{t\in[0,T]}$ denote the quadratic variation process and moreover let
$$\mathcal{E}(M)_t := \exp\left(M_t - \frac{1}{2}\langle M\rangle_t\right)\,.$$

We are given measurable functions
$$b : A\times[0,T]\times\mathbb{R}^d \to \mathbb{R}^d \text{ and } \sigma : [0,T]\times\mathbb{R}^d \to \mathbb{R}^{d\times d'}\,.$$
The state of the system is governed by the controlled SDE (1).

Assumption 2.1. The functions b and σ are continuous in t. There exists $K\ge 0$ and such that $\forall x,y\in\mathbb{R}^d, \forall a\in A, \forall t\in[0,T]$,
$$|b^a(t,x)-b^a(t,y)| + |\sigma(t,x)-\sigma(t,y)| \le K|x-y| \tag{9}$$
and
$$|\sigma(t,x)| \le K(1+|x|)\,,\ \ |b^a(t,x)| \le K(1+|x|+|a|)\,. \tag{10}$$
Under Assumption 2.1 we know that for any $(t,x)\in[0,T]\times\mathbb{R}^d$ and for any progressively measurable A-valued control process $\alpha=(\alpha_s)$ there is a unique strong solution to (1) which we denote $(X_s^{t,x,\alpha})_{s\in[t,T]}$. Let
$$f : A\times[0,T]\times\mathbb{R}^d \to \mathbb{R} \text{ and } g : \mathbb{R}^d\to\mathbb{R}$$
be two given measurable functions. Let us assume the following for the running gain function f and the terminal gain function g appearing in (2).

Assumption 2.2. There is a constant $K\ge 0$ such that $\forall x,y\in\mathbb{R}^d, \forall a\in A, \forall t\in[0,T]$
$$|g(x)-g(y)| + |f^a(t,x)-f^a(t,y)| \le K|x-y| \tag{11}$$
and
$$|f^a(t,x)| \le K(1+|x|+|a|)\,,\ \ |g(x)|\le K\,. \tag{12}$$
Under Assumption 2.2 the gain functional J given by (2) and the value function v given by (3) are well defined. Moreover, the value function v satisfies the Bellman equation (with derivatives existing almost everywhere, see Krylov [6, Chapter 4], or in the sense of viscosity solutions, see, e.g., Pham [15] or Fleming and Soner [11])
$$\begin{aligned}\partial_t v + \frac{1}{2}\mathrm{tr}(\sigma\sigma^\top D_x^2 v) + \sup_{a\in A}(b^a D_x v + f^a) = 0 \text{ on } [0,T)\times\mathbb{R}^n,\\ v(T,x) = g(x) \ \text{ on } x\in\mathbb{R}^d\,.\end{aligned} \tag{13}$$

Let us now state the additional assumptions required for our convergence result.

Assumption 2.3. Let us define for each fixed $(t,x,z)\in[0,T]\times\mathbb{R}^d\times\mathbb{R}^d$ the function
$$a(t,x,z) := \arg\max_{a\in A}\left(b^a(t,x)\sigma^{-1}(t,x)z + f^a(t,x)\right)\,. \tag{14}$$
We assume that the function $a(t,x,z)$ is measurable.

If the function $a \mapsto \big(b^a(t,x)\sigma^{-1}(t,x)z + f^a(t,x)\big)$ is convex for each fixed (t,x,z), which is in $[0,T]\times\mathbb{R}^d\times\mathbb{R}^d$, one can immediately see that Assumption 2.3 holds. More generally, this assumption can be verified using an appropriate measurable selection theorem. For example, if A is compact, then [7, Proposition D.5] shows that an appropriate measurable selection exists. If A is not compact but f is bounded, then [7, Proposition D.6] gives the same conclusion (using also that $z = D_x v(t,x)$ and Remark 2.8).

Assumption 2.4. There are constants $K, \theta \geq 0$ such that the following hold:

(1) (On the drift) For all $t \in [0,T]$, $x \in \mathbb{R}^d$, $a, a' \in A$,

$$|b^a(t,x) - b^{a'}(t,x)| \leq \sqrt{\theta}|a - a'| \tag{15}$$

and for all $t \in [0,T]$, $x \in \mathbb{R}^d$, $a \in A$ we have

$$|(b^a\sigma^{-1})(t,x)| < K\,. \tag{16}$$

(2) (On the control function) For all $t \in [0,T]$, $x, x', z, z' \in \mathbb{R}^d$, $a, a' \in A$ we have that

$$|a(t,x,z) - a(t,x,z')| \leq \sqrt{\theta}|z - z'|\,,$$
$$|a(t,x,z) - a(t,x',z)| \leq K|x - x'| \;\text{ and }\; |a(t,0,0)| \leq K\,.$$

(3) (On the running reward)

$$|f^a(t,x) - f^{a'}(t,x)| \leq \sqrt{\theta}|a - a'| \;\; \forall t \in [0,T], \forall x \in \mathbb{R}^d, \forall a, a' \in A\,.$$

Remark 2.5. Under Assumptions 2.2 and 2.4 we have that for all $t \in [0,T]$, $x, z, z' \in \mathbb{R}^d$ the following hold:

$$|f^{a(t,x,z)}(t,x) - f^{a(t,x,z')}(t,x)| \leq \theta|z - z'|$$

and

$$|f^{a(t,x,0)}(t,x)| \leq (K + K^2)(1 + |x|)\,.$$

Under Assumptions 2.1, 2.2, 2.3, and 2.4 there is an optimal control process and this fact will be used to prove the main results.

Remark 2.6. Due to results of Krylov [6] we know that (4) has a unique solution and moreover the map $[0,T]\times\mathbb{R}^d \ni (t,x) \mapsto D_x v(t,x) \in \mathbb{R}^d$ is bounded; see [6, Chapter 4, section 1, Theorem 1]. Hence, by Assumptions 2.3 and 2.4 we know that $(t,x) \mapsto a(t,x,\sigma(t,x)D_x v(t,x))$ is jointly measurable and Lipschitz in x. Thus, for each $(t,x) \in [0,T]\times\mathbb{R}^d$, the SDE

$$dX_s = b^{a(s,X_s,\sigma(s,X_s)D_x v(s,X_s))}(s,X_s)\,ds + \sigma(s,X_s)\,dW_s\,,\;\; s \in [t,T]\,, X_t = x$$

has a unique solution $X^{t,x}$. Then by the verification theorem, the process $\alpha^*_s := a(s,X_s,\sigma(s,X_s)D_x v(s,X_s))$ is the optimal control process for (3).

All the proofs will be completed in a new measure $\hat{\mathbb{P}}$ on $(\Omega,\mathcal{F})$ given in the following lemma. We will use $\hat{\mathbb{E}}$ to denote the expectation under the measure $\hat{\mathbb{P}}$.

Lemma 2.7. *Let Assumptions 2.1 and 2.2 together with* (16) *hold. Let* $(t,x) \in [0,T]\times\mathbb{R}^d$. *Let* $X = X^{t,x,\alpha^*}$ *be the solution to the SDE* (1) *started from* (t,x) *and controlled by the optimal control process* α^*. *Then* $d\hat{\mathbb{P}} := \mathcal{E}((b^{\alpha^*}\sigma^{-1})(\cdot,X)\bullet W)_T\,d\mathbb{P}$ *is a probability measure equivalent to* $\mathbb{P}$ *and the process*

$$\widehat{W}_s := W_s + \int_0^s b^{\alpha^*_r}(r,X_r)\sigma^{-1}(r,X_r)\,dr$$

is a $\hat{\mathbb{P}}$*-Wiener process.*

Proof. This is an immediate consequence of (16) and Girsanov's theorem. □

Remark 2.8. From Krylov [6, Chapter 4, section 1, Theorem 1] we get that there is a constant $C > 0$ such that for all $(t,x) \in [0,T)\times\mathbb{R}^d$ we have that $|D_x v(t,x)| \leq C$.

3. Convergence of gradient iteration algorithm

The following theorem gives the convergence result for Algorithm 2.

Theorem 3.1. *Let Assumptions 2.1, 2.2, 2.3, and 2.4 hold. Let v be the solution to (4) and let $(v^n)_{n\in\mathbb{N}}$ be the approximation sequence given by Algorithm 2. Then there is $q \in (0,1)$ depending only on K, θ, T and the initial guess $v^0 = v^0(t,x)$ such that for all $(t,x) \in [0,T]\times\mathbb{R}^d$ there exists $C = C(t,x)$ such that*

$$|v(t,x) - v^n(t,x)|^2 \le C(t,x)q^n\,.$$

The main idea of the proof consists of noticing that Algorithm 2 can be seen as an iteration on the level of BSDEs. Using Lemma A.2 we see that on the level of BSDEs this iteration is contractive. Finally we need to use known results on the connection between BSDEs and solutions to the HJB equation.

Proof of Theorem 3.1. We prove the main result in several steps. First, we show how to rewrite the gradient iteration algorithm as an iteration on the level of BSDEs. On the nth step of the algorithm we need to solve the linear PDE with Lipschitz continuous coefficients (8). Let v^n be the solution to (8) and recall that

$$a^n(t,x) = \arg\max_{a\in A}\big((b^a D_x v^{n-1})(t,x) + f^a(t,x)\big) = a(t,x,\sigma(t,x)D_x v^{n-1}(t,x))\,.$$

Since we are working with the linear PDE with Lipschitz continuous coefficients, we have v^n in $C^{1,2}([0,T)\times\mathbb{R}^d)$. Let $X = X^{t,x,\alpha^*}$ be the solution to the SDE (1) started from (t,x) and controlled by the optimal control process α^*; see Remark 2.6. From Itô's formula we then get that

$$\begin{aligned} dv^n(s,X_s) &= \left[\partial_t v^n(s,X_s) + \frac{1}{2}\mathrm{tr}(\sigma\sigma^\top D_x^2 v^n)(s,X_s) + (b^{\alpha_s^*}D_x v^n)(s,X_s)\right] ds \\ &\quad + (D_x v^n \sigma)(s,X_s)\, dW_s \\ &= \left[(b^{\alpha_s^*}D_x v^n)(s,X_s) - (b^{a^n}D_x v^{n-1})(s,X_s) - f^{a^n}(s,X_s)\right] ds \\ &\quad + (D_x v^n\sigma)(s,X_s)\, dW_s\,. \end{aligned}$$

Let

$$F_s(z) := b^{a(s,X_s,z)}(s,X_s)\sigma^{-1}(s,X_s)z + f^{a(s,X_s,z)}(s,X_s)$$

and

$$Y_t^n := v^n(t,X_t)\,,\quad Z_t^n := \sigma(t,X_t)D_x v^n(t,X_t)\,,\ \ \xi := g(X_T). \tag{17}$$

Then we may write

$$Y_t^n = \xi - \int_t^T \left[(b^{\alpha_s^*}\sigma^{-1})(s,X_s)Z_s^n - F_s(Z_s^{n-1})\right] ds - \int_t^T Z_s^n\, dW_s\,. \tag{18}$$

Let $\hat{\mathbb{P}}$ and $\widehat{W}$ be given by Lemma 2.7. Hence (18) becomes

$$Y_t^n = \xi + \int_t^T F_s(Z_s^{n-1})\, ds - \int_t^T Z_s^n\, d\widehat{W}_s\,. \tag{19}$$

Consider now the following BSDE:

$$\begin{aligned} Y_{t'}^{t,x} &= \xi + \int_{t'}^T b^{a(s,X_s^{t,x},Z_s^{t,x})}(s,X_s^{t,x})\sigma^{-1}(s,X_s^{t,x})Z_s^{t,x} + f^{a(s,X_s^{t,x},Z_s^{t,x})}(s,X_s^{t,x})\, ds \\ &\quad - \int_{t'}^T Z_s^{t,x}\, d\widehat{W}_s,\ \ t' \in [t,T]\,, \end{aligned} \tag{20}$$

where the superscript means that the forward process started from (t,x). Hence, we can define

$$w(t,x) := Y_t^{t,x} \ \text{ and } \ \sigma(t,x)D_x w(t,x) := Z_t^{t,x}. \tag{21}$$

Therefore by (14) we have

$$b^{a(t,x,z)}(t,x)\sigma^{-1}(t,x)z + f^{a(t,x,z)}(t,x) = \max_{a\in A}\left(b^a(t,x)\sigma^{-1}(t,x)z + f^a(t,x)\right).$$

Thus, by Pham [15, Theorem 6.3.3], the function $w = w(t,x)$ solves the HJB equation (4). Notice that here is the crucial point where the fact that we use the optimal control α^* plays a role. Indeed with other control processes we couldn't claim that w solves the HJB equation. By uniqueness of the viscosity solution to the HJB equation (see the strong comparison principle from [15, Theorem 4.4.5]), we can conclude that $w = v$ and therefore w is the value function of our stochastic control problem. Therefore, the BSDE (20) is the BSDE corresponding to the value function. Notice that (20) is a quadratic BSDE, since in the generator we have a product of two Lipschitz functions which depend on Z. The existence of the solution to (20) under our assumptions can be obtained by applying Theorem A.9 in the case when the terminal cost is bounded for our stochastic control problem.

Using Remark 2.8, the fact that $\sigma^{-1}(s,X_s)Z_s = D_x v(s,X_s)$, and Assumption 2.4, for all $s \in [t,T]$ we get that

$$\begin{aligned}
&|F_s(Z_s) - F_s(Z_s^{n-1})| \\
&\quad\le |\sigma^{-1}(s,X_s)Z_s||b^{a(s,X_s,Z_s)}(s,X_s) - b^{a(s,X_s,Z_s^{n-1})}(s,X_s)| \\
&\qquad + |b^{a(s,X_s,Z_s^{n-1})}(s,X_s)\sigma^{-1}(s,X_s)||Z_s - Z_s^{n-1}| + \theta|Z_s - Z_s^{n-1}| \\
&\quad\le (C\theta + K + \theta)|Z_s - Z_s^{n-1}|\,.
\end{aligned} \tag{22}$$

Moreover, recalling $\xi = g(X_T)$, we get that $\xi \in L^2(\Omega, \mathcal{F}_T, \hat{\mathbb{P}})$ from the higher moment estimates for the solution of the SDE and from the Lipschitz property of g. Similarly $F_s(0) \in \hat{\mathbb{H}}^2$ by Assumption 2.4. We may thus apply Lemma A.2 and hence, due to (19) and (22), we have $q \in (0,1)$ and $\gamma \ge 0$ such that for all $t \in [0,T]$

$$e^{\gamma t}\hat{\mathbb{E}}|Y_t - Y_t^n|^2 + \|Z - Z^n\|^2_{\hat{\mathbb{H}}^2_\gamma} \le q\|Z - Z^{n-1}\|^2_{\hat{\mathbb{H}}^2_\gamma}\,. \tag{23}$$

Therefore, from (21) and (17), we have $v(t,x) = Y_t^{t,x}$ and $v^n(t,x) = Y_t^{n,t,x}$ and by (23) we obtain

$$\begin{aligned}
e^{\gamma t}|v(t,x) - v^n(t,x)|^2 &\le \hat{\mathbb{E}}[e^{\gamma t}|Y_t^{t,x} - Y_t^{n,t,x}|^2] + \|Z^{t,x} - Z^{n,t,x}\|^2_{\hat{\mathbb{H}}^2_\gamma} \\
&\le q^n\|Z^{t,x} - Z^{0,t,x}\|^2_{\hat{\mathbb{H}}^2_\gamma}\,.
\end{aligned}$$

Hence

$$\begin{aligned}
&|v(t,x) - v^n(t,x)|^2 \\
&\quad\le q^n\hat{\mathbb{E}}\int_t^T e^{\gamma(T-t)}|\sigma(s,X_s^{t,x,\alpha^\star})|^2|D_x v(s,X_s^{t,x,\alpha^\star}) - D_x v^0(s,X_s^{t,x,\alpha^\star})|^2\,ds\,.
\end{aligned}$$

This finishes the proof. □

4. Convergence of policy improvement

Theorem 4.1. *Let Assumptions 2.1, 2.2, 2.3, and 2.4 hold. Let v be the solution to (4) and let $(v^n)_{n\in\mathbb{N}}$ be the approximation sequence given by Algorithm 1. Then there is $q \in (0,1)$ depending only on K, θ, T and the initial guess $v^0 = v^0(t,x)$ such that for all $(t,x) \in [0,T]\times\mathbb{R}^d$ there exists $C = C(t,x)$ such that*

$$|v(t,x) - v^n(t,x)|^2 \le C(t,x)q^n\,.$$

The proof of Theorem 4.1 is similar to that of Theorem 3.1 except that the iteration on the level of BSDEs is nonstandard.

Proof of Theorem 4.1. Let v^n be the solution to (5) and recall that

$$a^n(t,x) = \arg\max_{a\in A}\left((b^a D_x v^{n-1})(t,x) + f^a(t,x)\right) = a(t,x,\sigma(t,x)D_x v^{n-1}(t,x))\,.$$

As before, let $X = X^{t,x,\alpha^*}$ be the solution to the SDE (1) started from (t,x) and controlled by the optimal control process α^*; see Remark 2.6. By Itô's formula

$$\begin{aligned} dv^n(s,X_s) &= \left[\partial_t v^n(s,X_s) + \frac{1}{2}\mathrm{tr}(\sigma\sigma^\top D_x^2 v^n)(s,X_s) + (b^{\alpha_s^*} D_x v^n)(s,X_s)\right] ds \\ &\quad + (D_x v^n \sigma)(s,X_s)\, dW_s \\ &= \left[(b^{\alpha_s^*} D_x v^n)(s,X_s) - (b^{a^n} D_x v^n)(s,X_s) - f^{a^n}(s,X_s)\right] ds \\ &\quad + (D_x v^n \sigma)(s,X_s)\, dW_s\,. \end{aligned}$$

Let

$$F_s(z,Z) := b^{a(s,X_s,z)}(s,X_s)\sigma^{-1}(s,X_s)Z + f^{a(s,X_s,z)}(s,X_s)\,.$$

Recalling that the control α^* and the associated diffusion X are fixed we can write

$$Y_t^n = \xi - \int_t^T \left[(b^{\alpha_s^*}\sigma^{-1})(s,X_s)Z_s^n - F_s(Z_s^{n-1}, Z_s^n)\right] ds - \int_t^T Z_s^n dW_s. \tag{24}$$

Let $\hat{\mathbb{P}}$ and $\widehat{W}$ be given by Lemma 2.7. Then (24) becomes

$$Y_t^n = \xi + \int_t^T F_s(Z_s^{n-1}, Z_s^n) ds - \int_t^T Z_s^n d\widehat{W}_s\,. \tag{25}$$

Similarly as in Theorem 3.1, consider the BSDE

$$\begin{aligned} Y_{t'}^{t,x} &= \xi + \int_{t'}^T b^{a(s,X_s^{t,x},Z_s^{t,x})}(s,X_s^{t,x})\sigma^{-1}(s,X_s^{t,x})Z_s^{t,x} + f^{a(s,X_s^{t,x},Z_s^{t,x})}(s,X_s^{t,x})\, ds \\ &\quad - \int_{t'}^T Z_s^{t,x}\, d\widehat{W}_s,\ \ t' \in [t,T]\,. \end{aligned} \tag{26}$$

In same way we can show that $v(t,x) = Y_t^{t,x}$ is the value function of our stochastic control problem. As before, from Krylov [6, Chapter 4, section 1, Theorem 1], we get that there is a constant $C > 0$ such that for all $(t,x) \in [0,T)\times\mathbb{R}^d$ we have that $|D_x v(t,x)| \le C$. Moreover, as before, using Remark 2.8, the fact that $\sigma^{-1}(s,X_s)Z_s = D_x v(s,X_s)$, and Assumption 2.4, for all $s \in [t,T]$ we get that

$$\begin{aligned} &|F_s(Z_s,Z_s) - F_s(Z_s^{n-1},Z_s^n)| \\ &\quad\le \theta|\sigma^{-1}(s,X_s)Z_s||Z_s - Z_s^{n-1}| \\ &\qquad + |b^{a(s,X_s,Z_s^{n-1})}(s,X_s)\sigma^{-1}(s,X_s)||Z_s - Z_s^n| + \theta|Z_s - Z_s^{n-1}| \\ &\quad\le \theta C|Z_s - Z_s^{n-1}| + K|Z_s - Z_s^n| + \theta|Z_s - Z_s^{n-1}|\,. \end{aligned} \tag{27}$$

Finally we note that $\xi \in L^2(\Omega,\mathcal{F}_T,\hat{\mathbb{P}})$ and $F_s(0,0) \in \hat{\mathbb{H}}^2$, so by Lemma A.5, together with (25) and (26), we have $q \in (0,1)$ and $\gamma \ge 0$ such that for all $t \in [0,T]$

$$e^{\gamma t}\hat{\mathbb{E}}|Y_t - Y_t^n|^2 + \|Z - Z^n\|^2_{\hat{\mathbb{H}}^2_\gamma} \le q\|Z - Z^{n-1}\|^2_{\hat{\mathbb{H}}^2_\gamma}\,. \tag{28}$$

Similarly as before, using (28), we conclude that

$$\begin{aligned} &|v(t,x) - v^n(t,x)|^2 \\ &\quad\le q^n\hat{\mathbb{E}}\int_t^T e^{\gamma(T-t)}|\sigma(s,X_s^{t,x,\alpha^*})|^2|D_x v(s,X_s^{t,x,\alpha^*}) - D_x v^0(s,X_s^{t,x,\alpha^*})|^2\, ds\,. \end{aligned}$$

This concludes the proof of the theorem. □

Remark 4.2. Consider briefly the situation where the diffusion coefficient also depends on the control, i.e., $\sigma = \sigma^a(t,x)$. After applying Itô's formula to $v^n(s,X_s)$ and substituting the solution to the linear PDE for v^n we get

$$\begin{aligned} dv^n(s,X_s) =& \big[(b^{\alpha_s^*} D_x v^n)(s,X_s) + \frac{1}{2}\mathrm{tr}(\sigma^{\alpha_s^*}(\sigma^{\alpha_s^*})^\top D_x^2 v^n)(s,X_s) \\ & - (b^{a^n} D_x v^n)(s,X_s) - \frac{1}{2}\mathrm{tr}(\sigma^{a^n}(\sigma^{a^n})^\top D_x^2 v^n)(s,X_s) - f^{a^n}(s,X_s)\big]\, ds \\ & + (D_x v^n \sigma)(s,X_s)\, dW_s\,. \end{aligned}$$

The resulting object can be seen as a second order BSDE (2BSDE). Analysis of 2BSDEs goes beyond the scope of this paper.

Remark 4.3. Let us briefly consider the infinite-time-horizon control problem. In this case we consider a constant $\lambda > 0$ and the gain functional:

$$J(x,\alpha) = \mathbb{E}\left[\int_0^\infty e^{-\lambda s} f^\alpha(X_s^x)\, ds\right].$$

It is known that the Bellman PDE for the value function is

$$\lambda v - \frac{1}{2}\mathrm{tr}(\sigma\sigma^\top D_x^2 v) - \sup_{a\in A}(b^a D_x v + f^a) = 0 \ \text{ on } \ \mathbb{R}^d\,.$$

The linear PDE from the iteration of the policy improvement algorithm then is

$$\lambda v^n - \frac{1}{2}\mathrm{tr}(\sigma\sigma^\top D_x^2 v^n) - b^{a^n} D_x v^n - f^{a^n} = 0 \ \text{ on } \ \mathbb{R}^d\,,$$

where

$$a^n(x) = \arg\max_{a\in A}\left(b^a(x) D_x v^{n-1}(x) + f^a(x)\right).$$

After applying Itô's formula we get

$$\begin{aligned} dv^n(X_s) &= \left[\frac{1}{2}\mathrm{tr}(\sigma\sigma^\top D_x^2 v^n)(X_s) + (b^{\alpha_s^*} D_x v^n)(X_s)\right] ds + (D_x v^n \sigma)(X_s)\, dW_s \\ &= \big[(b^{\alpha_s^*} D_x v^n)(X_s) - (b^{a^n} D_x v^n)(X_s) - f^{a^n}(X_s) + \lambda v^n(X_s)\big]\, ds \\ &\quad + (D_x v^n \sigma)(X_s)\, dW_s\,. \end{aligned}$$

Let $Y_t^n := v^n(X_t)$ and $Z_t^n := \sigma(X_t) D_x v^n(X_t)$. Then after change of measure we may write

$$dY_s^n = -\big[(b^{a(X_s, Z_s^{n-1})}\sigma^{-1})(X_s) Z_s^n + f^{a(X_s,Z_s^{n-1})}(X_s) - \lambda Y_s^n\big]\, ds + Z_s^n\, d\widehat{W}_s\,. \quad (29)$$

Let

$$F_s(z,Z) := (b^{a(X_s,z)}\sigma^{-1})(X_s) Z + f^{a(X_s,z)}(X_s)\,.$$

Hence (29) becomes

$$dY_s^n = \big[-F_s(Z_s^{n-1}, Z_s^n) + \lambda Y_s^n\big]\, ds + Z_s^n\, d\widehat{W}_s\,, \ \ s\in[0,\infty)\,. \quad (30)$$

To proceed, we need a suitable contraction-type inequality for this infinite time horizon BSDE. Buckdahn and Peng [8] studied infinite time horizon BSDEs and have proved existence and uniqueness of their solutions for sufficiently large values of λ. To get the required contraction-type inequality we can use similar calculations as in Fuhrman and Tessitore [10, Theorems 3.2 and 3.7], where they use Banach's fixed point theorem to show existence and uniqueness of solutions to infinite time horizon BSDEs. Hence, for sufficiently large $\lambda > 0$ we would obtain results analogous to Theorem 4.1 as well as the other theorems in the article.

5. Policy improvement

We want to show that the policy obtained at each step of Algorithm 1 is an improvement on the one from the previous step. This is formulated as Theorem 5.1 below. Note that we do not require Assumption 2.4 here.

Theorem 5.1. *Let Assumptions 2.1, 2.2, and 2.3 hold. Assume that there exists $K \geq 0$ such that $\forall t \in [0,T]$, $\forall x \in \mathbb{R}^d$, and $\forall a \in A$*

$$|b^a \sigma^{-1}(t,x)| < K.$$

Fix $n \in \mathbb{N}$. Let v^n and v^{n+1} be the solutions of (5) at steps n and $n+1$ of the algorithm. Then for all $t \in [0,T]$, $x \in \mathbb{R}^d$ it holds that

$$v^{n+1}(t,x) \geq v^n(t,x).$$

Proof. Let $X = X^{t,x,\alpha^*}$ be the solution to the SDE (1) started from (t,x) and controlled by the optimal control process α^*; see Remark 2.6. Then, as in the proof of Theorem 4.1, we get that for $k = n, n+1$ with $Y^k = Y^{k,t,x} = v^k(\cdot, X^{t,x,\alpha^*})$ and with $Z^k = Z^{k,t,x} = (\sigma D_x v^k)(\cdot, X^{t,x,\alpha^*})$ we have the BSDE representation

$$Y_t^k = \xi + \int_t^T F_s(Z_s^{k-1}, Z_s^k) ds - \int_t^T Z_s^k \, d\widehat{W}_s \,, \;\; k = n, n+1,$$

where

$$F_s(z,Z) := b^{a(s,X_s,z)}(s,X_s)\sigma^{-1}(s,X_s)Z + f^{a(s,X_s,z)}(s,X_s)\,.$$

Let us denote for $s \in [t,T]$ and $z \in \mathbb{R}^d$

$$\phi_s^2(z) := F_s(Z_s^n, z) \;\text{ and }\; \phi_s^1(z) := F_s(Z_s^{n-1}, z)\,.$$

Hence, notice that by the definition of the a^{n+1} (see (6)), we have for all $s \in [t,T]$ that

$$\begin{aligned}
\phi_s^2(Z_s^n) =& F_s(Z_s^n, Z_s^n) = b^{a^{n+1}}(s,X_s)\sigma^{-1}(s,X_s)Z_s^n + f^{a^{n+1}}(s,X_s) \\
=& \max_{a \in A} \left((b^a D_x v^n)(s,X_s) + f^a(s,X_s) \right) \geq (b^{a^n} D_x v^n)(s,X_s) + f^{a^n}(s,X_s) \\
=& F_s(Z_s^{n-1}, Z_s^n) = \phi_s^1(Z_s^n).
\end{aligned}$$

Therefore by the comparison principle for BSDEs (see Lemma A.6), we get

$$Y_{t'}^{n+1} \geq Y_{t'}^n \;\; \forall t' \in [t,T].$$

Hence, we have

$$v^{n+1}(t,x) = Y_t^{n+1,t,x} \geq Y_t^{n,t,x} = v^n(t,x)\,.$$

□

Remark 5.2. It is perhaps interesting to note that the comparison principle for BSDEs cannot be used to deduce that in the gradient iteration algorithm we have an "improvement" at each step. Indeed, let us write the BSDE representation of the two steps of gradient iteration for $n, n+1 \in \mathbb{N}$,

$$Y_t^n = \xi + \int_t^T F_s(Z_s^{n-1}) ds - \int_t^T Z_s^n \, d\widehat{W}_s$$

and

$$Y_t^{n+1} = \xi + \int_t^T F_s(Z_s^n) ds - \int_t^T Z_s^{n+1} \, d\widehat{W}_s \,,$$

where

$$F_s(z) := b^{a(s,X_s,z)}(s,X_s)\sigma^{-1}(s,X_s)z + f^{a(s,X_s,z)}(s,X_s).$$

In order to apply a comparison principle for BSDEs (see Lemma A.6), we would need to have $F_s(Z_s^{n-1}) \leq F_s(Z_s^n)$. Nevertheless we observe that

$$\begin{aligned} F_s(Z_s^{n-1}) &= b^{a(s,X_s,Z_s^{n-1})}\sigma^{-1}(s,X_s)Z_s^{n-1} + f^{a(s,X_s,Z_s^{n-1})} \\ &= b^{a^n}(s,X_s)\sigma^{-1}(s,X_s)Z_s^{n-1} + f^{a^n}(s,X_s) \\ &= \max_{a\in A}(b^a(s,X_s)D_x v^{n-1}(s,X_s) + f^a(s,X_s)). \end{aligned}$$

Similarly,

$$\begin{aligned} F_s(Z_s^{n}) &= b^{a(s,X_s,Z_s^{n})}\sigma^{-1}(s,X_s)Z_s^{n} + f^{a(s,X_s,Z_s^{n})} \\ &= b^{a^{n+1}}(s,X_s)\sigma^{-1}(s,X_s)Z_s^{n} + f^{a^{n+1}}(s,X_s) \\ &= \max_{a\in A}(b^a(s,X_s)D_x v^{n}(s,X_s) + f^a(s,X_s)). \end{aligned}$$

From the above calculations we have no way to conclude that $F_s(Z_s^{n-1}) \leq F_s(Z_s^n)$. Thus the gradient iteration algorithm is not guaranteed to be improving the policy with each step.

6. Stability under Perturbations to Solution of the Linear PDE

In this section we study a stability property of the policy improvement algorithm under perturbations to solutions of the linear PDE (5) since in practical applications one will only solve this equation approximately. Of course the maximization step (6) of Algorithm 1 can now be performed only with this approximate solution, thus feeding the errors into further iterations.

Let ε be a parameter (or a set of parameters), which determines the accuracy of our approximation to the solution of the linear PDE (5). Let π_ε^n be the policy at iteration n obtained from an approximate solution to the linear PDE. Let v_ε^n denote the solution to

$$\begin{aligned} \partial_t v_\varepsilon^n + \frac{1}{2}\mathrm{tr}(\sigma\sigma^\top D_x^2 v_\varepsilon^n) + b^{\pi_\varepsilon^n} D_x v_\varepsilon^n + f^{\pi_\varepsilon^n} = 0 \text{ on } [0,T)\times\mathbb{R}^d\,, \\ v_\varepsilon^n(T,\cdot) = g \ \text{ on } x\in\mathbb{R}^d\,. \end{aligned} \tag{31}$$

At step n of Algorithm 1 we approximate the solution to the equation above (this is PDE (5) but with π_ε^n replacing a^n everywhere). We will denote such approximation by $\tilde{v}_\varepsilon^n$. The policy function for the next iteration step is then given by

$$\pi_\varepsilon^{n+1}(t,x) = a(t,x,(\sigma D_x\tilde{v}_\varepsilon^n)(t,x)) = \arg\max_{a\in A}[(b^a D_x\tilde{v}_\varepsilon^n)(t,x) + f^a(t,x)]\,,$$

recalling that the function $a = a(t,x,z)$ was defined in (14). We need to assume that $(t,x)\mapsto D_x\tilde{v}_\varepsilon^n$ is bounded so that π_ε^{n+1} is Lipschitz in x so that the solution to (31) is $C^{1,2}([0,T]\times\mathbb{R}^d)$. This assumption is not really a restriction as we know that the gradient of the value function is bounded under our assumptions; see Krylov [6, Chapter 4, section 1, Theorem 1] and also Remark 2.6. Any reasonable approximation should retain this property.

Theorem 6.1. *Let Assumptions 2.1, 2.2, 2.3, and 2.4 hold. Let $(v^n)_{n\in\mathbb{N}}$ be the approximation sequence given by Algorithm 1. Let $(v_\varepsilon^n)_{n\in\mathbb{N}}$ be the approximation sequence given by (31). Let α^* and X^{t,x,α^*} be the optimal control process for (3) and the associated diffusion started from $(t,x)\in[0,T]\times\mathbb{R}^d$. Assume that $D_x\tilde{v}_\varepsilon^n$ is uniformly bounded. Define*

$$E_{t,x}^k := \left\| \left[(\sigma(D_x v_\varepsilon^k - D_x\tilde{v}_\varepsilon^k))(\cdot, X^{t,x,\alpha^*})\right]\mathcal{E}^{-1/2}((b^{\alpha^*}\sigma^{-1})(\cdot,X^{t,x,\alpha^*})\bullet W)_T \right\|_{\mathbb{H}^2}.$$

Then there is $q \in (0,1)$ and $\gamma > 0$, depending only on K, θ, T, such that for all $(t,x) \in [0,T] \times \mathbb{R}^d$ there exists $C = C(t,x)$ such that

$$|v^n(t,x) - v^n_\varepsilon(t,x)|^2 \le C(t,x)q^n + 2e^{\gamma(T-t)} \sum_{k=1}^{n} q^k E^{n-k}\,.$$

Proof. Let $X = X^{t,x,\alpha^*}$ be the solution to the SDE (1) started from (t,x) and controlled by the optimal control process α^*; see Remark 2.6. By applying Itô's formula to v^n_ε we get

$$\begin{aligned} dv^n_\varepsilon(s,X_s) &= \left[\partial_t v^n_\varepsilon(s,X_s) + \frac{1}{2}\mathrm{tr}(\sigma\sigma^\top D^2_x v^n_\varepsilon)(s,X_s) + (b^{\alpha^*_s} D_x v^n_\varepsilon)(s,X_s)\right] ds \\ &\quad + (D_x v^n_\varepsilon \sigma)(s,X_s)\, dW_s \\ &= \left[(b^{\alpha^*_s} D_x v^n_\varepsilon)(s,X_s) - \left(b^{\pi^n_\varepsilon} D_x v^n_\varepsilon + f^{\pi^n_\varepsilon}\right)(s,X_s)\right] ds \\ &\quad + (D_x v^n_\varepsilon \sigma)(s,X_s)\, dW_s. \end{aligned}$$

Let us denote

$$Y^n_{t,\varepsilon} := v^n_\varepsilon(t,X_t)\,, \quad Z^n_{t,\varepsilon} := \sigma(t,X_t) D_x v^n_\varepsilon(t,X_t)\,, \quad \xi := g(X_T)\,,$$

$$F_s(z,Z) := b^{a(s,X_s,z)}(s,X_s)\sigma^{-1}(s,X_s)Z + f^{a(s,X_s,z)}(s,X_s)$$

and

$$\tilde{Z}^{n-1}_{t,\varepsilon} := \sigma(t,X_t) D_x \tilde{v}^{n-1}_\varepsilon(t,X_t),$$

where $\tilde{v}^{n-1}_\varepsilon$ is an approximate solution to corresponding PDE. Then using this notation, we may write

$$Y^n_{t,\varepsilon} = \xi - \int_t^T \left[(b^{\alpha^*_s}\sigma^{-1})(s,X_s) Z^n_{s,\varepsilon} - F_s(\tilde{Z}^{n-1}_{s,\varepsilon}, Z^n_{s,\varepsilon})\right] ds - \int_t^T Z^n_{s,\varepsilon}\, dW_s\,.$$

Let $\hat{\mathbb{P}}$ and $\widehat{W}$ be given by Lemma 2.7. Then the above equation becomes

$$Y^n_{t,\varepsilon} = \xi + \int_t^T F_s(\tilde{Z}^{n-1}_{s,\varepsilon}, Z^n_{s,\varepsilon}) ds - \int_t^T Z^n_{s,\varepsilon}\, d\widehat{W}_s\,.$$

We want to study the difference of $(Y^n_\varepsilon, Z^n_\varepsilon)$ with (Y^n, Z^n), where (Y^n, Z^n) solves the BSDE (25).

$$\begin{aligned} \hat{\mathbb{E}}[e^{\gamma t}|Y^n_t - Y^n_{t,\varepsilon}|^2] + \|Z^n - Z^n_\varepsilon\|^2_{\hat{\mathbb{H}}^2_\gamma} &\le 2\hat{\mathbb{E}}[e^{\gamma t}|Y_t - Y^n_t|^2] + 2\|Z - Z^n\|^2_{\hat{\mathbb{H}}^2_\gamma} \\ &\quad + 2\hat{\mathbb{E}}[e^{\gamma t}|Y_t - Y^n_{t,\varepsilon}|^2] + 2\|Z - Z^n_\varepsilon\|^2_{\hat{\mathbb{H}}^2_\gamma}\,, \end{aligned} \tag{32}$$

where (Y,Z) solves the BSDE (26). Due to (27) and

$$|F_s(Z_s,Z_s) - F_s(\tilde{Z}^{n-1}_{s,\varepsilon}, Z^n_{s,\varepsilon})| \le (C\theta+\theta)|Z_s - \tilde{Z}^{n-1}_{s,\varepsilon}| + K|Z_s - Z^n_{s,\varepsilon}|\,,$$

we can apply Lemma A.5. Hence, there is $\tilde{q} \in (0,1/2)$ and $\gamma > 0$ such that

$$\hat{\mathbb{E}}[e^{\gamma t}|Y_t - Y^n_t|^2] + \|Z - Z^n\|^2_{\hat{\mathbb{H}}^2_\gamma} \le \tilde{q}\|Z - Z^{n-1}\|^2_{\hat{\mathbb{H}}^2_\gamma}$$

and

$$\hat{\mathbb{E}}[e^{\gamma t}|Y_t - Y^n_{t,\varepsilon}|^2] + \|Z - Z^n_\varepsilon\|^2_{\hat{\mathbb{H}}^2_\gamma} \le \tilde{q}\|Z - \tilde{Z}^{n-1}_\varepsilon\|^2_{\hat{\mathbb{H}}^2_\gamma}\,.$$

Therefore we continue the estimate (32)

$$\begin{aligned}
&\hat{\mathbb{E}}[e^{\gamma t}|Y_t^n - Y_{t,\varepsilon}^n|^2] + \|Z^n - Z_\varepsilon^n\|^2_{\hat{\mathbb{H}}^2_\gamma} \\
&\leq 2\hat{\mathbb{E}}[e^{\gamma t}|Y_t - Y_t^n|^2] + 2\|Z - Z^n\|^2_{\hat{\mathbb{H}}^2_\gamma} + 2\hat{\mathbb{E}}[e^{\gamma t}|Y_t - Y_{t,\varepsilon}^n|^2] + 2\|Z - Z_\varepsilon^n\|^2_{\hat{\mathbb{H}}^2_\gamma} \\
&\leq 2\tilde{q}\|Z - Z^{n-1}\|^2_{\hat{\mathbb{H}}^2_\gamma} + 2\tilde{q}\|Z - \tilde{Z}_\varepsilon^{n-1}\|^2_{\hat{\mathbb{H}}^2_\gamma} \\
&\leq 2\tilde{q}\|Z - Z^{n-1}\|^2_{\hat{\mathbb{H}}^2_\gamma} + 4\tilde{q}\|Z - Z_\varepsilon^{n-1}\|^2_{\hat{\mathbb{H}}^2_\gamma} + 4\tilde{q}\|Z_\varepsilon^{n-1} - \tilde{Z}_\varepsilon^{n-1}\|^2_{\hat{\mathbb{H}}^2_\gamma} \\
&\leq 2\tilde{q}^2\|Z - Z^{n-2}\|^2_{\hat{\mathbb{H}}^2_\gamma} + 4\tilde{q}^2\|Z - \tilde{Z}_\varepsilon^{n-2}\|^2_{\hat{\mathbb{H}}^2_\gamma} + 4\tilde{q}\|Z_\varepsilon^{n-1} - \tilde{Z}_\varepsilon^{n-1}\|^2_{\hat{\mathbb{H}}^2_\gamma} \\
&\leq 2\tilde{q}^2\|Z - Z^{n-2}\|^2_{\hat{\mathbb{H}}^2_\gamma} + 8\tilde{q}^2\|Z - Z_\varepsilon^{n-2}\|^2_{\hat{\mathbb{H}}^2_\gamma} + 8\tilde{q}^2\|Z_\varepsilon^{n-2} - \tilde{Z}_\varepsilon^{n-2}\|^2_{\hat{\mathbb{H}}^2_\gamma} \\
&\qquad + 4\tilde{q}\|Z_\varepsilon^{n-1} - \tilde{Z}_\varepsilon^{n-1}\|^2_{\hat{\mathbb{H}}^2_\gamma} \leq \cdots \\
&\leq 2\tilde{q}^n\|Z - Z^0\|^2_{\hat{\mathbb{H}}^2_\gamma} + 2^{n+1}\tilde{q}^n\|Z - Z_\varepsilon^0\|^2_{\hat{\mathbb{H}}^2_\gamma} + \sum_{k=1}^{n} 2^{k+1}\tilde{q}^k\|Z_\varepsilon^{n-k} - \tilde{Z}_\varepsilon^{n-k}\|^2_{\hat{\mathbb{H}}^2_\gamma}\,.
\end{aligned}$$

This concludes the proof of the theorem. □

7. Stability under Perturbation of the Maximization

In this section we study a stability property of the gradient iteration algorithm under perturbations to maximization procedure (7). Let $\bar{v}^n$ be the solution to corresponding PDE at iteration n of the gradient iteration algorithm, where instead of obtaining the control function corresponding to the exact maximum

$$a^n(t,x) = a(t,x,(\sigma D_x \bar{v}^n)(t,x)) = \arg\max_{a\in A}((b^a D_x \bar{v}^n + f^a)(t,x))$$

we only solve this maximization problem approximately and so we are dealing with a control function of the form

$$\bar{a}(t,x,(\sigma D_x \bar{v}^n)(t,x)) := a(t,x,(\sigma D_x \bar{v}^n)(t,x)) + \varepsilon(t,x,(\sigma D_x \bar{v}^n)(t,x))\,,$$

where the function $\varepsilon = \varepsilon(t,x,z)$ determines the accuracy of our approximation.

Theorem 7.1. *Let Assumptions 2.1, 2.2, 2.3, and 2.4 hold. Let $(v^n)_{n\in\mathbb{N}}$ be the approximation sequence given by Algorithm 2. Let $(\bar{v}^n)_{n\in\mathbb{N}}$ be the approximation sequence given by the perturbations to the maximization procedure and assume that $v^0 = \bar{v}^0$. Let α^* and X^{t,x,α^*} be the optimal control process for (3) and the associated diffusion started from $(t,x) \in [0,T]\times\mathbb{R}^d$. Define*

$$E_{t,x}^{k+1} := \left\| \left[1 + |D_x \bar{v}^k(\cdot, X^{t,x,\alpha^*})|\right] \mathcal{E}^{-1/2}((b^{\alpha^*}\sigma^{-1})(\cdot, X^{t,x,\alpha^*}) \bullet W)_T \right\|^2_{\mathbb{H}^2}\,,$$

$$\varepsilon^{k+1} = \sup_{(s,y)\in[t,T]\times\mathbb{R}^d} |\varepsilon(s,y,(\sigma D_x \bar{v}^k)(s,y))|^2\,.$$

Then there is $q \in (0,1)$ and $\gamma > 0$, depending only on K, θ, T, such that for all $(t,x) \in [0,T]\times\mathbb{R}^d$ there exists $C = C(t,x)$ such that

$$|\bar{v}^n(t,x) - v^n(t,x)|^2 \leq C(t,x)q^n + e^{\gamma(T-t)}\frac{2\theta}{C\theta + K + \theta}\sum_{k=1}^{n} q^{n-k+1}\varepsilon^k E_{t,x}^k\,.$$

Proof. Let $X = X^{t,x,\alpha^*}$ be the solution to the SDE (1) started from (t,x) and controlled by the optimal control process α^*; see Remark 2.6. As in the proof of Theorem 3.1 we can write two BSDEs we get after the change of measure given by Lemma 2.7. The first BSDE arises from the perturbations of the maximization:

$$\bar{Y}_t^n = \xi + \int_t^T \bar{F}_s(\bar{Z}_s^{n-1})\,ds - \int_t^T \bar{Z}_s^n\,d\widehat{W}_s\,,$$

where

$$\bar{F}_s(z) = b^{\bar{a}(s,X_s,z)}\sigma^{-1}(s,X_s)z + f^{\bar{a}(s,X_s,z)}(s,X_s)\,.$$

The second BSDE arises from the gradient iteration algorithm with the maximization performed exactly:

$$Y_t^n = \xi + \int_t^T F_s(Z_s^{n-1})\,ds - \int_t^T Z_s^n\,d\widehat{W}_s\,,$$

where

$$F_s(z) = b^{a(s,X_s,z)}\sigma^{-1}(s,X_s)z + f^{a(s,X_s,z)}(s,X_s)\,.$$

We want to study the difference of $(\bar{Y}^n, \bar{Z}^n)$ with (Y^n, Z^n). Hence, notice that

$$\begin{aligned} e^{\gamma t}\hat{\mathbb{E}}|\bar{Y}_t^n - Y_t^n|^2 + \|\bar{Z}^n - Z^n\|^2_{\hat{\mathbb{H}}^2_\gamma} \le 2e^{\gamma t}\hat{\mathbb{E}}|Y_t - Y_t^n|^2 + 2\|Z - Z^n\|^2_{\hat{\mathbb{H}}^2_\gamma} \\ + 2e^{\gamma t}\hat{\mathbb{E}}|Y_t - \bar{Y}_t^n|^2 + 2\|Z - \bar{Z}^n\|^2_{\hat{\mathbb{H}}^2_\gamma}\,, \end{aligned} \tag{33}$$

where (Y, Z) solves (20). Therefore, since

$$|F_s(Z_s) - F_s(\bar{Z}_s^{n-1})| \le (C\theta + K + \theta)|Z_s - \bar{Z}_s^{n-1}|\,,$$

we can apply Lemma A.7 so that there is $q \in (0,1)$ and $\gamma > 0$ such that

$$\begin{aligned} e^{\gamma t}\hat{\mathbb{E}}|Y_t - \bar{Y}_t^n|^2 + \|Z - \bar{Z}^n\|^2_{\hat{\mathbb{H}}^2_\gamma} \le & q\|Z - \bar{Z}^{n-1}\|^2_{\hat{\mathbb{H}}^2_\gamma} \\ & + \frac{q}{C\theta + K + \theta}\|\bar{F}(\bar{Z}^{n-1}) - F(\bar{Z}^{n-1})\|^2_{\hat{\mathbb{H}}^2_\gamma}\,. \end{aligned} \tag{34}$$

Now we need to estimate the second term of the right-hand side (RHS). Notice that by Assumption 2.4 the following holds:

$$\begin{aligned} &|\bar{F}_s(\bar{Z}_s^{n-1}) - F_s(\bar{Z}_s^{n-1})| \\ &\le |\sigma^{-1}(s,X_s)\bar{Z}_s^{n-1}||b^{\bar{a}(s,X_s,\bar{Z}_s^{n-1})}(s,X_s) - b^{a(s,X_s,\bar{Z}_s^{n-1})}(s,X_s)| \\ &\quad + |f^{\bar{a}(s,X_s,\bar{Z}_s^{n-1})}(s,X_s) - f^{a(s,X_s,\bar{Z}_s^{n-1})}(s,X_s)| \\ &\le \sqrt{\theta}|\sigma^{-1}(s,X_s)\bar{Z}_s^{n-1}||\varepsilon(s,X_s,\bar{Z}_s^{n-1})| + \sqrt{\theta}|\varepsilon(s,X_s,\bar{Z}_s^{n-1})|\,. \end{aligned} \tag{35}$$

Hence by (35) we have

$$\|\bar{F}(\bar{Z}^{n-1}) - F(\bar{Z}^{n-1})\|^2_{\hat{\mathbb{H}}^2_\gamma} \le \theta\|(1 + |\sigma^{-1}(\cdot,X)\bar{Z}^{n-1}|)\varepsilon(\cdot,X,\bar{Z}^{n-1})\|^2_{\hat{\mathbb{H}}^2_\gamma}\,. \tag{36}$$

By inequalities (33), (34), (36), and the result of Theorem 3.1 and since $\bar{Y}_t^{t,x,n} = \bar{v}^n(t,x)$, $Y_t^{t,x,n} = v^n(t,x)$ as well as $Z^0 = \bar{Z}^0$, we conclude that

$$\begin{aligned} &e^{\gamma t}|\bar{v}^n(t,x) - v^n(t,x)|^2 \\ &\quad \le 2e^{\gamma t}\hat{\mathbb{E}}|Y_t - Y_t^n|^2 + \|Z - Z^n\|^2_{\hat{\mathbb{H}}^2_\gamma} + 2e^{\gamma t}\hat{\mathbb{E}}|Y_t - \bar{Y}_t^n|^2 + \|Z - \bar{Z}^n\|^2_{\hat{\mathbb{H}}^2_\gamma} \\ &\quad \le 2q\|Z - Z^{n-1}\|^2_{\hat{\mathbb{H}}^2_\gamma} + 2q\|Z - \bar{Z}^{n-1}\|^2_{\hat{\mathbb{H}}^2_\gamma} + \frac{2q}{C\theta + K + \theta}\|\bar{F}(\bar{Z}^{n-1}) - F(\bar{Z}^{n-1})\|^2_{\hat{\mathbb{H}}^2_\gamma} \\ &\quad \le 4q^n\|Z - Z^0\|^2_{\hat{\mathbb{H}}^2_\gamma} + \sum_{k=1}^n q^k\frac{2\theta}{C\theta + K + \theta}\|(1 + |\sigma^{-1}(\cdot,X)\bar{Z}^{n-k}|)\varepsilon(\cdot,X,\bar{Z}^{n-k})\|^2_{\hat{\mathbb{H}}^2_\gamma}\,. \end{aligned}$$

□

We obtain the same result for the policy improvement algorithm.

Theorem 7.2. *Let Assumptions 2.1, 2.2, 2.3, and 2.4 hold. Let $(v^n)_{n\in\mathbb{N}}$ be the approximation sequence given by Algorithm 1. Let $(\bar{v}^n)_{n\in\mathbb{N}}$ be the approximation sequence given by the perturbations to the maximization procedure. Let α^* and*

X^{t,x,α^*} be the optimal control process for (3) and the associated diffusion started from $(t,x) \in [0,T] \times \mathbb{R}^d$. Define

$$E^k_{t,x} := \left\| \left[1 + D_x \bar{v}^k(\cdot, X^{t,x,\alpha^*}) \right] \mathcal{E}^{-1/2}((b^{\alpha^*}\sigma^{-1})(\cdot, X^{t,x,\alpha^*}) \bullet W)_T \right\|^2_{\mathbb{H}^2} ,$$

$$\varepsilon^{k+1} = \sup_{(s,y)\in[t,T]\times\mathbb{R}^d} |\varepsilon(s, y, (\sigma D_x \bar{v}^k)(s,y))|^2 .$$

Then there is $q \in (0,1)$ and $\gamma > 0$, depending only on K, θ, T, such that for all $(t,x) \in [0,T] \times \mathbb{R}^d$ there is $C = C(t,x)$ such that

$$|\bar{v}^n(t,x) - v^n(t,x)|^2 \le C(t,x) q^n + \frac{2\theta}{\max(C\theta + \theta, K)} e^{\gamma(T-t)} \sum_{k=1}^{n} q^k \varepsilon^{n-k+1} E^{n-k+1}_{t,x} .$$

Proof. Let $X = X^{t,x,\alpha^*}$ be the solution to the SDE (1) started from (t,x) and controlled by the optimal control process α^*; see Remark 2.6. Due to Theorem 4.1 we can write two BSDEs we get after the change of measure: first from the perturbation and second from the gradient iteration

$$\bar{Y}^n_t = \xi + \int_t^T \bar{F}_s(\bar{Z}^{n-1}_s, \bar{Z}^n_s)\, ds - \int_t^T \bar{Z}^n_s \, d\widehat{W}_s ,$$

$$Y^n_t = \xi + \int_t^T F_s(Z^{n-1}_s, Z^n_s)\, ds - \int_t^T Z^n_s \, d\widehat{W}_s ,$$

where

$$\bar{F}_s(z,Z) := b^{\bar{a}(s,X_s,z)}(s,X_s)\sigma^{-1}(s,X_s)Z + f^{\bar{a}(s,X_s,z)}(s,X_s) ,$$
$$F_s(z,Z) := b^{a(s,X_s,z)}(s,X_s)\sigma^{-1}(s,X_s)Z + f^{a(s,X_s,z)}(s,X_s) .$$

Similarly, we want to study the difference of $(\bar{Y}^n, \bar{Z}^n)$ with (Y^n, Z^n). Hence, notice that

$$\begin{aligned} e^{\gamma t}\mathbb{E}|\bar{Y}^n_t - Y^n_t|^2 + \|\bar{Z}^n - Z^n\|^2_{\hat{\mathbb{H}}^2_\gamma} &\le 2e^{\gamma t}\mathbb{E}|Y_t - Y^n_t|^2 + \|Z - Z^n\|^2_{\hat{\mathbb{H}}^2_\gamma} \\ &\quad + 2e^{\gamma t}\mathbb{E}|Y_t - \bar{Y}^n_t|^2 + \|Z - \bar{Z}^n\|^2_{\hat{\mathbb{H}}^2_\gamma} , \end{aligned} \tag{37}$$

where (Y,Z) solves (4.1). Therefore, since

$$|F_s(Z_s, Z_s) - F_s(\bar{Z}^{n-1}_s, \bar{Z}^n_s)| \le \theta C|Z_s - \bar{Z}^{n-1}_s| + K|Z_s - \bar{Z}^n_s| + \theta|Z_s - \bar{Z}^{n-1}_s| ,$$

we can apply Lemma A.8 so that there is $q \in (0,1)$ and $\gamma > 0$ such that

$$\begin{aligned} e^{\gamma t}\mathbb{E}|Y_t - \bar{Y}^n_t|^2 + \|Z - \bar{Z}^n\|^2_{\hat{\mathbb{H}}^2_\gamma} &\le q\|Z - \bar{Z}^{n-1}\|^2_{\hat{\mathbb{H}}^2_\gamma} \\ &\quad + \frac{q}{\max(C\theta + \theta, K)} \|\bar{F}(\bar{Z}^{n-1}, \bar{Z}^n) - F(\bar{Z}^{n-1}, \bar{Z}^n)\|^2_{\hat{\mathbb{H}}^2_\gamma} . \end{aligned} \tag{38}$$

Now we need to estimate the second term of the RHS. Notice that by Assumption 2.4 we have that

$$\begin{aligned} &|\bar{F}_s(\bar{Z}^{n-1}_s, \bar{Z}^n_s) - F_s(\bar{Z}^{n-1}_s, \bar{Z}^n_s)| \\ &\le \left| b^{\bar{a}(s,X_s,\bar{Z}^{n-1}_s)}(s,X_s)\sigma^{-1}(s,X_s)\bar{Z}^n_s - b^{a(s,X_s,\bar{Z}^{n-1}_s)}(s,X_s)\sigma^{-1}(s,X_s)\bar{Z}^n_s \right| \\ &\qquad + \left| f^{\bar{a}(s,X_s,\bar{Z}^{n-1}_s)}(s,X_s) - f^{a(s,X_s,\bar{Z}^{n-1}_s)}(s,X_s) \right| \\ &\le \sqrt{\theta}|\sigma^{-1}(s,X_s)\bar{Z}^n_s||\varepsilon(s,X_s,\bar{Z}^{n-1}_s)| + \sqrt{\theta}|\varepsilon(s,X_s,\bar{Z}^{n-1}_s)| . \end{aligned} \tag{39}$$

Hence by (39) we have

$$\|\bar{F}(\bar{Z}^{n-1}, \bar{Z}^n) - F(\bar{Z}^{n-1}, \bar{Z}^n)\|^2_{\hat{\mathbb{H}}^2_\gamma} \le \theta \|(1 + |\sigma^{-1}(s,X)\bar{Z}^n|)\varepsilon(\cdot, X, \bar{Z}^{n-1})\|^2_{\hat{\mathbb{H}}^2_\gamma} . \tag{40}$$

By inequalities (37), (38), (40), by the result of Theorem 4.1, and by $\bar{Y}_t^{t,x,n} = \bar{v}^n(t,x)$, $Y_t^{t,x,n} = v^n(t,x)$ we conclude that

$$\begin{aligned}
&e^{\gamma t}|\bar{v}^n(t,x) - v^n(t,x)|^2 \\
&\quad \le 2e^{\gamma t}\hat{\mathbb{E}}|Y_t - Y_t^n|^2 + \|Z - Z^n\|^2_{\hat{\mathbb{H}}^2_\gamma} + 2e^{\gamma t}\hat{\mathbb{E}}|Y_t - \bar{Y}_t^n|^2 + \|Z - \bar{Z}^n\|^2_{\hat{\mathbb{H}}^2_\gamma} \\
&\quad \le 2q\|Z - Z^{n-1}\|^2_{\hat{\mathbb{H}}^2_\gamma} + 2q\|Z - \bar{Z}^{n-1}\|^2_{\hat{\mathbb{H}}^2_\gamma} \\
&\qquad + \frac{2q}{\max(C\theta + \theta, K)}\|\bar{F}(\bar{Z}^{n-1}) - F(\bar{Z}^{n-1})\|^2_{\hat{\mathbb{H}}^2_\gamma} \\
&\quad \le 2q^n\|Z - Z^0\|^2_{\hat{\mathbb{H}}^2_\gamma} + 2q^n\|Z - \bar{Z}^0\|^2_{\hat{\mathbb{H}}^2_\gamma} \\
&\qquad + \sum_{k=1}^{n} q^k \frac{2\theta}{\max(C\theta + \theta, K)}\|(1 + |\sigma^{-1}(\cdot, X)\bar{Z}^{n-k+1}|)\varepsilon(\cdot, X, \bar{Z}^{n-k})\|^2_{\hat{\mathbb{H}}^2_\gamma}\,.
\end{aligned}$$

□

8. Example

In this section we would like to consider an example when Assumptions 2.1, 2.2, 2.3, and 2.4 hold. Let $t \mapsto s(t)$ and $t \mapsto k(t)$ be continuous functions for $t \in [0,T]$. Consider the state which is governed by the controlled SDE

$$dX_t = s(t)\sin\alpha_t\,dt + \sqrt{2}\,dW_t\,,\ \ t \in [0,T]\,,$$

and consider the cost functional

$$J(t,x,\alpha) = \mathbb{E}\left[\int_t^T k(s)\cos\alpha_s\,ds + g(X_T)\right].$$

The aim is to maximize J over admissible controls $\alpha \in \mathcal{A}$. The value function $v = \sup_{\alpha\in\mathcal{A}} J(t,x,\alpha)$ satisfies the Bellman PDE

$$\partial_t v + D_x^2 v + \sup_{a\in A}[s(t)\sin a\, D_x v + k(t)\cos a] = 0\,,\ \ \text{on } [0,T)\times\mathbb{R}\,,$$

with the terminal condition $v(T,x) = g(x) := \arctan(x)$. Hence, the optimal control is

$$a(t,x) = \arctan\left(\frac{s(t)D_x v}{k(t)}\right).$$

It is easy to check that Assumptions 2.1, 2.2, 2.3, and 2.4 hold for this problem. Therefore, the Bellman PDE becomes

$$\partial_t v + D_x^2 v + \frac{\frac{(s(t)D_x v)^2}{k(t)}}{\sqrt{1 + \left(\frac{s(t)D_x v}{k(t)}\right)^2}} + \frac{k(t)}{\sqrt{1 + \left(\frac{s(t)D_x v}{k(t)}\right)^2}} = 0\,. \tag{41}$$

We can solve this problem using the policy improvement algorithm by approximating the Bellman PDE with a sequence of linear PDEs:

Step 1. Make an initial choice of control $a^0(t,x)$.

Step 2. For $n = 0, 1, \ldots$:

- Evaluation step: Find a solution $v^n = v^n(t,x)$ to the linear PDE

$$\partial_t v^n + D_x^2 v^n + s(t)\sin a^n D_x v^n + \cos a^n = 0\,. \tag{42}$$

- Improvement step: Find a new policy $a^{n+1} = a^{n+1}(t,x)$ such that

$$a^{n+1}(t,x) = \arctan\left(\frac{s(t)D_x v^n}{k(t)}\right).$$

Step 3. Iterate the process until no changes occur in the controls updates.

One can do similar calculations in the case of the gradient iteration algorithm.

We will solve (41) and (42) by the finite difference method. For simplicity, let us choose $s(t) = 1$ and $k(t) = 1$ for all $t \in [0, T]$. In Figure 8.1, one can see the logarithm of the error between the value function obtained by the iterative methods, by the policy improvement algorithm, and by the gradient iteration algorithm at every step and the value function obtained by the solution of the Bellman PDE. This shows the fast convergence of the policy improvement method for our example in one dimension. In Figure 8.2, we can see that after only a few steps the policies obtained from the policy improvement algorithm are close to the exact one. Finally, in Figure 8.3, we plot the value function and the policy from the solution of the Bellman PDE.

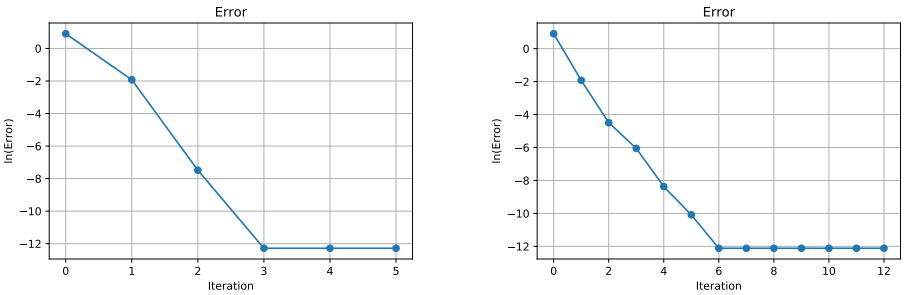

FIGURE 8.1. Plot for the logarithm of the error between the value function obtained by the iterative methods, by the policy improvement, and by the gradient iteration algorithms (from left to right) at every step and the value function obtained by the solution of the Bellman PDE. Note that the convergence stops once we have reached the accuracy of the finite-difference solver.

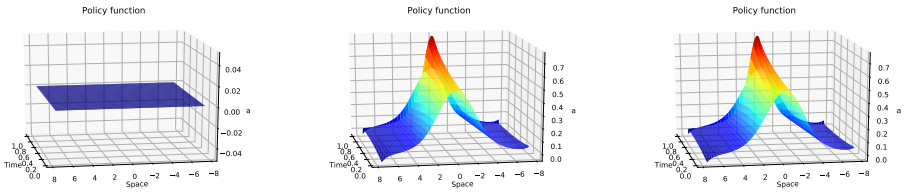

FIGURE 8.2. Plot of the initial policy and policies obtained by the policy improvement algorithm at Steps 1 and 5

APPENDIX A. SOME RESULTS FROM THEORY OF BSDES

We fix a finite horizon $T \in (0, \infty)$. We fix a filtered probability space $(\Omega, \mathcal{F}, \mathbb{F} = (\mathcal{F}_t)_{0 \le t \le T}, \mathbb{P})$. Let there be a d'-dimensional Wiener martingale on this space.

Lemma A.1. *Let $F : \Omega \times [0, T] \times \mathbb{R}^d \to \mathbb{R}$ be a measurable function that satisfies the following conditions: The process $(F_t(0))_{t \in [0,T]}$ is in $\mathbb{H}^2$. Moreover there is a constant $\theta > 0$ such that*

$$|F_t(z)| \le |F_t(0)| + \theta |z| \ \ \forall z \in \mathbb{R}^d, t \in [0, T], \ a.s.$$

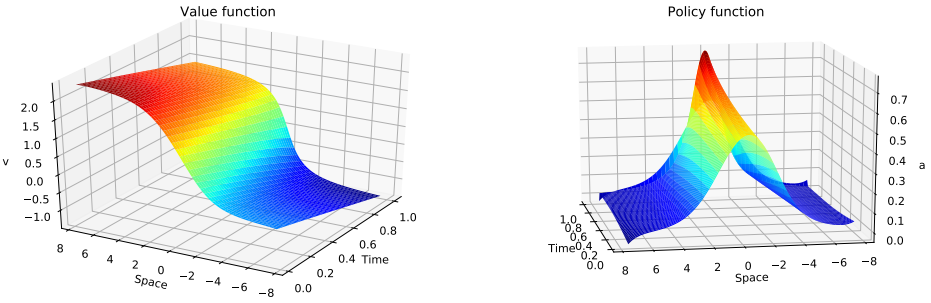

FIGURE 8.3. Plot of the true value function and the policy

Then, for every $\xi \in L^2(\Omega, \mathcal{F}_T)$ *and* $z \in \mathbb{H}^2$*, there is a unique solution* $(Y, Z) \in \mathcal{S}^2 \times \mathbb{H}^2$ *to*

$$Y_t = \xi + \int_t^T F_s(z_s)\, ds - \int_t^T Z_s\, dW_s,\ t \in [0,T],\ \mathbb{P}\text{-}a.s. \tag{43}$$

Proof. This follows immediately from, e.g., Pham [15, Theorem 6.2.1]. □

Lemma A.2. *Let* $F : \Omega \times [0,T] \times \mathbb{R}^d \to \mathbb{R}$ *satisfy the hypothesis of Lemma A.1. Fix* $\xi \in L^2(\Omega, \mathcal{F}_T)$*. Let* $\Phi : \mathbb{H}^2 \ni z \mapsto (Y, Z) \in \mathcal{S}^2 \times \mathbb{H}^2$*, where* (Y, Z) *is the unique solution to* (43)*. Moreover assume that for* $z^1, z^2 \in \mathbb{H}^2$ *the following condition satisfies that there is a constant* $\theta > 0$ *such that*

$$|F_t(z_t^1) - F_t(z_t^2)| \le \theta |z_t^1 - z_t^2|\,,\ \ t \in [0,T],\ a.s. \tag{44}$$

Then there is $\gamma > 0$ *and* $q \in (0,1)$ *such that for* $(Y^i, Z^i) := \Phi(z^i)$*,* $i = 1, 2$*, and any* $t \in [0,T]$ *we have*

$$\mathbb{E}\left[e^{\gamma t}|Y_t^1 - Y_t^2|^2\right] + \|Z^1 - Z^2\|^2_{\mathbb{H}^2_\gamma} \le q\|z^1 - z^2\|^2_{\mathbb{H}^2_\gamma}\,.$$

The proof is well known and is included, e.g., as part of Pham [15, Proof of Theorem 6.2.1]. We provide it here for the convenience of the reader and before we proceed we need to make the following observation.

Remark A.3. Assume that $Y \in \mathcal{S}^2$ and $Z \in \mathbb{H}^2$ and let

$$M_t := \int_0^t e^{\gamma s} Z_s\, Y_s\, dW_s\,.$$

Then $\sup_{t \le T} |M_t| \in L^1(\Omega, \mathcal{F}_T)$ and hence M_t is a uniformly integrable martingale. Indeed, from the Burkholder–Davis–Gundy inequality and the Young inequality we get

$$\begin{aligned}
\mathbb{E}\left[\sup_{t\le T}|M_t|\right] &\le C_1\mathbb{E}\left[\left(\int_0^T e^{2\gamma s}|Y_s|^2|Z_s|^2 ds\right)^{1/2}\right] \\
&\le e^{\gamma T}C_1\mathbb{E}\left[\left(\sup_{t\le T}|Y_t|^2\int_0^T |Z_s|^2 ds\right)^{1/2}\right] \\
&\le \frac{e^{\gamma T}}{2}C_1\mathbb{E}\left[\sup_{t\le T}|Y_t|^2 + \int_0^T |Z_s|^2 ds\right] < \infty.
\end{aligned}$$

Proof of Lemma A.2. Consider $\gamma > 0$ which we will fix later. We denote $\delta z := z^1 - z^2$, $\delta Z := Z^1 - Z^2$, $\delta Y := Y^1 - Y^2$, and $\delta F := F(z^1) - F(z^2)$. We then apply

Itô's formula to $e^{\gamma t}|\delta Y_t|^2$:

$$e^{\gamma t}|\delta Y_t|^2 + \int_t^T e^{\gamma s}|\delta Z_s|^2 ds = \int_t^T e^{\gamma s}(2\delta Y_s\, \delta F_s - \gamma|\delta Y_s|^2)ds \\ - 2\int_t^T e^{\gamma s}\delta Z_s\, \delta Y_s\, dW_s\,.$$

Due to Remark A.3, the stochastic integral vanishes by taking expectation. Hence

$$\mathbb{E}\left[e^{\gamma t}|\delta Y_t|^2 + \int_t^T e^{\gamma s}|\delta Z_s|^2 ds\right] = \mathbb{E}\left[\int_t^T e^{\gamma s}(2\delta Y_s\, \delta F_s - \gamma|\delta Y_s|^2)\, ds\right].$$

By the Lipschitz property of the generator and by the Young inequality we continue our estimate, noting that for any $\varepsilon > 0$, we have

$$\mathbb{E}\left[e^{\gamma t}|\delta Y_t|^2 + \int_t^T e^{\gamma s}|\delta Z_s|^2\, ds\right] \le \mathbb{E}\left[\int_t^T e^{\gamma s}(2\theta|\delta Y_s||\delta z_s| - \gamma|\delta Y_s|^2)ds\right] \\ \le \mathbb{E}\left[\int_t^T e^{\gamma s}\left(\theta(\varepsilon|\delta Y_s|^2 + \varepsilon^{-1}|\delta z_s|^2) - \gamma|\delta Y_s|^2\right) ds\right].$$

Choose ε such that $\gamma = \varepsilon\theta$. Thus

$$\mathbb{E}\left[e^{\gamma t}|\delta Y_t|^2 + \int_t^T e^{\gamma s}|\delta Z_s|^2 ds\right] \le \mathbb{E}\left[\int_t^T e^{\gamma s}\left(\theta\varepsilon^{-1}|\delta z_s|^2\right) ds\right] \le \frac{\theta^2}{\gamma}\|\delta z\|^2_{\mathbb{H}^2_\gamma}. \quad (45)$$

Hence, from (45) we have that for $\gamma > \theta^2$ and any $t \in [0,T]$

$$\mathbb{E}\left[e^{\gamma t}|Y_t^1 - Y_t^2|^2\right] + \|Z^1 - Z^2\|^2_{\mathbb{H}^2_\gamma} \le q\|z^1 - z^2\|^2_{\mathbb{H}^2_\gamma},$$

where $q = \frac{\theta^2}{\gamma} \in (0,1)$. This concludes the proof of the lemma. □

Lemma A.4. *Let $F : \Omega \times [0,T] \times \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ be a measurable function such that the process $(F_t(0,0))_{t\in[0,T]}$ is in $\mathbb{H}^2$ and such that there are $\theta, K > 0$ so that for all $t \in [0,T]$, $z, Z \in \mathbb{R}^d$ we have*

$$|F_t(z,Z)| \le |F_t(0,0)| + \theta|z| + K|Z| \ \ a.s.$$

If $\xi \in L^2(\Omega, \mathcal{F}_T)$ and $z \in \mathbb{H}^2$, then there is a unique solution $(Y,Z) \in \mathcal{S}^2 \times \mathbb{H}^2$ to

$$Y_t = \xi + \int_t^T F_s(z_s, Z_s)\, ds - \int_t^T Z_s\, dW_s\,, \ \ t \in [0,T]\,, \ \mathbb{P}\text{-}a.s. \quad (46)$$

Proof. This follows immediately from, e.g., Pham [15, Theorem 6.2.1]. □

Lemma A.5. *Let $F : \Omega\times[0,T]\times\mathbb{R}^d\times\mathbb{R}^d \to \mathbb{R}$ satisfy the hypothesis of Lemma A.4. Fix $\xi \in L^2(\Omega, \mathcal{F}_T)$. Let $\Phi : \mathbb{H}^2 \ni z \mapsto (Y,Z) \in \mathcal{S}^2 \times \mathbb{H}^2$, where (Y,Z) is the unique solution to (46). Moreover assume that for $z^1, z^2 \in \mathbb{H}^2$ the following condition satisfies that there are constants $\theta, K > 0$ such that*

$$|F_t(z_t^1, Z_t^1) - F_t(z_t^2, Z_t^2)| \le \theta|z_t^1 - z_t^2| + K|Z_t^1 - Z_t^2|\,, \ \ t \in [0,T]\,, \ a.s.,$$

where $(Y^i, Z^i) := \Phi(z^i)$, $i = 1,2$. Then there is $\gamma > 0$ and $q \in (0,1)$ such that for any $t \in [0,T]$ we have

$$\mathbb{E}\left[e^{\gamma t}|Y_t^1 - Y_t^2|^2\right] + \|Z^1 - Z^2\|^2_{\mathbb{H}^2_\gamma} \le q\|z^1 - z^2\|^2_{\mathbb{H}^2_\gamma}.$$

Moreover, there is $\gamma > 0$ such that $q \in (0, 1/2)$.

Proof. Consider $\gamma > 0$ which we will fix later. We denote $\delta z := z^1 - z^2$, $\delta Z := Z^1 - Z^2$, and $\delta Y := Y^1 - Y^2$. We then apply Itô's formula to $e^{\gamma t}|\delta Y_t|^2$:

$$\begin{aligned} e^{\gamma t}|\delta Y_t|^2 &+ \int_t^T e^{\gamma s}|\delta Z_s|^2 ds \\ &= \int_t^T e^{\gamma s}\left(2\delta Y_s(F_s(z_s^1, Z_s^1) - F_s(z_s^2, Z_s^2)) - \gamma|\delta Y_s|^2\right) ds \\ &\quad - 2\int_t^T e^{\gamma s}\delta Z_s \delta Y_s \, dW_s \,. \end{aligned}$$

The expectation of the stochastic integral is 0 due to Remark A.3. Hence, by taking expectation we derive from the equality above that

$$\begin{aligned} \mathbb{E}\left[e^{\gamma t}|\delta Y_t|^2 + \int_t^T e^{\gamma s}|\delta Z_s|^2 ds\right] \\ = \mathbb{E}\int_t^T e^{\gamma s}\left(2\delta Y_s(F_s(z_s^1, Z_s^1) - F_s(z_s^2, Z_s^2)) - \gamma|\delta Y_s|^2\right) ds \,. \end{aligned}$$

By the Lipschitz property of the generator and by the Young inequality we observe that, for any $\varepsilon > 0$,

$$\begin{aligned} \mathbb{E}\left[e^{\gamma t}|\delta Y_t|^2 + \int_t^T e^{\gamma s}|\delta Z_s|^2 ds\right] &\le \mathbb{E}\int_t^T e^{\gamma s}\left(2\delta Y_s(\theta|\delta z_s| + K|\delta Z_s|) - \gamma|\delta Y_s|^2\right) ds \\ &\le \mathbb{E}\int_t^T e^{\gamma s}\left((\theta+K)\varepsilon|\delta Y_s|^2 + \theta\varepsilon^{-1}|\delta z_s|^2 + K\varepsilon^{-1}|\delta Z_s|^2 - \gamma|\delta Y_s|^2\right) ds \,. \end{aligned}$$

Take $\gamma > 0$ sufficiently large so that $\tilde{q} := \max(\frac{(\theta+K)K}{\gamma}, \frac{(\theta+K)\theta}{\gamma}) \in (0, 1/2)$. Choose ε such that $\gamma = (\theta + K)\varepsilon$. Thus

$$\mathbb{E}\left[e^{\gamma t}|\delta Y_t|^2 + (1-\tilde{q})\int_t^T e^{\gamma s}|\delta Z_s|^2 ds\right] \le \mathbb{E}\left[\int_t^T e^{\gamma s}\tilde{q}|\delta z_s|^2 ds\right]. \tag{47}$$

Dividing by $1 - \tilde{q} \in (1/2, 1)$ we obtain

$$\mathbb{E}\left[e^{\gamma t}|\delta Y_t|^2 + \int_t^T e^{\gamma s}|\delta Z_s|^2 ds\right] \le q\mathbb{E}\left[\int_t^T e^{\gamma s}|\delta z_s|^2 ds\right],$$

where $q := \frac{\tilde{q}}{1-\tilde{q}}$. Since $0 < \tilde{q} < 1/2$ we have that $q \in (0, 1)$. Therefore from (47) we have for any $t \in [0, T]$

$$\mathbb{E}\left[e^{\gamma t}|Y_t^1 - Y_t^2|^2\right] + \|Z^1 - Z^2\|^2_{\mathbb{H}^2_\gamma} \le q\|z^1 - z^2\|^2_{\mathbb{H}^2_\gamma} \,.$$

By choosing γ such that $\tilde{q} \in (0, 1/3)$ we get that $q \in (0, 1/2)$. This concludes the proof of the lemma. □

We now state a comparison principle for BSDEs.

Lemma A.6. *Consider the following BSDEs:*

$$Y_t^i = \xi^i + \int_t^T \phi^i(s, Z_s^i)ds - \int_t^T Z_s^i \, dW_s \,, \; t \in [0, T], \;\; i = 1, 2.$$

Assume that $\xi^i \in L^2(\Omega, \mathcal{F}_T)$, $i = 1, 2$, and $\xi^1 \le \xi^2$ a.s. Let $\phi^i : \Omega \times [0, T] \times \mathbb{R}^d \to \mathbb{R}$, $i = 1, 2$, be such that for all $z \in \mathbb{R}^d$ the processes $(\phi^i(t, z))_{t\in[0,T]}$ are progressively measurable, $\phi^i(t, 0) \in \mathbb{H}^2$, and such that there is $\theta > 0$ so that for all $t \in [0, T]$, $z, z' \in \mathbb{R}^d$ we have

$$|\phi^i(t, z) - \phi^i(t, z')| \le \theta|z - z'| \text{ a.s.}$$

Moreover, suppose that for $Z^1, Z^2 \in \mathbb{H}^2$ it holds that

$$\phi^1(t, Z_t^1) \leq \phi^2(t, Z_t^1) \ \ \forall t \in [0,T]\,, \ a.s.$$

Then $Y_t^1 \leq Y_t^2$ for all $0 \leq t \leq T$ a.s.

Proof. This follows from, e.g., Pham [15, Theorem 6.2.2]. □

The following two lemmas are auxiliary results we need in Section 7.

Lemma A.7. *Let $F, \bar{F} : \Omega \times [0,T] \times \mathbb{R}^d \to \mathbb{R}$ be measurable functions and let F satisfy the hypotheses of Lemmas A.1 and A.2. Fix $\xi \in L^2(\Omega, \mathcal{F}_T)$. Let $\bar{z}, z, Z, \bar{Z} \in \mathbb{H}^2$ and $Y, \bar{Y} \in \mathcal{S}^2$ be such that*

$$\bar{Y}_t = \xi + \int_t^T \bar{F}_s(\bar{z}_s)\, ds - \int_t^T \bar{Z}_s\, dW_s\,, \ \ t \in [0,T]\,,$$

and

$$Y_t = \xi + \int_t^T F_s(z_s)\, ds - \int_t^T Z_s\, dW_s\,, \ \ t \in [0,T]\,.$$

Then there is $\gamma > 0$ and $q \in (0,1)$ such that for $t \in [0,T]$ we have

$$e^{\gamma t}\mathbb{E}|\bar{Y}_t - Y_t|^2 + \|\bar{Z} - Z\|^2_{\mathbb{H}^2_\gamma} \leq q\|\bar{z} - z\|^2_{\mathbb{H}^2_\gamma} + \frac{q}{\theta}\|\bar{F}(\bar{z}) - F(\bar{z})\|^2_{\mathbb{H}^2_\gamma}.$$

Proof. Consider $\gamma > 0$ which we will fix later. We denote $\tilde{Y} := \bar{Y} - Y$, $\tilde{Z} := \bar{Z} - Z$, and $\tilde{z} := \bar{z} - z$. We then apply Itô's formula to $e^{\gamma t}|\tilde{Y}_t|^2$:

$$\begin{aligned} e^{\gamma t}|\tilde{Y}_t|^2 + \int_t^T e^{\gamma s}|\tilde{Z}_s|^2\, ds &= \int_t^T e^{\gamma s}(2\tilde{Y}_s\,(\bar{F}_s(\bar{z}_s) - F_s(z_s)) - \gamma|\tilde{Y}_s|^2)\, ds \\ &\quad - 2\int_t^T e^{\gamma s}\tilde{Z}_s\,\tilde{Y}_s\, dW_s\,. \end{aligned}$$

Due to Remark A.3, the stochastic integral vanishes by taking expectation. Hence

$$\mathbb{E}\left[e^{\gamma t}|\tilde{Y}_t|^2 + \int_t^T e^{\gamma s}|\tilde{Z}_s|^2\, ds\right] = \mathbb{E}\left[\int_t^T e^{\gamma s}(2\tilde{Y}_s\,(\bar{F}_s(\bar{z}_s) - F_s(z_s)) - \gamma|\tilde{Y}_s|^2)\, ds\right]. \tag{48}$$

Notice that due to (44) for all $s \in [t,T]$ it holds that

$$|\bar{F}_s(\bar{z}_s) - F_s(z_s)| \leq |\bar{F}_s(\bar{z}_s) - F_s(\bar{z}_s)| + |F_s(\bar{z}_s) - F_s(z_s)| \leq |\bar{F}_s(\bar{z}_s) - F_s(\bar{z}_s)| + \theta|\bar{z}_s - z_s|.$$

Then by the Young inequality we continue our estimate (48), noting that for any $\delta > 0$, we have

$$\begin{aligned} &\mathbb{E}\left[e^{\gamma t}|\tilde{Y}_t|^2 + \int_t^T e^{\gamma s}|\tilde{Z}_s|^2\, ds\right] \\ &\quad \leq \mathbb{E}\left[\int_t^T e^{\gamma s}(2|\tilde{Y}_s|(|\bar{F}_s(\bar{z}_s) - F_s(\bar{z}_s)| + \theta|\bar{z}_s - z_s|) - \gamma|\tilde{Y}_s|^2)ds\right] \\ &\quad \leq \mathbb{E}\left[\int_t^T e^{\gamma s}\left((1+\theta)\delta|\tilde{Y}_s|^2 + \delta^{-1}(\theta|\tilde{z}_s|^2 + |\bar{F}_s(\bar{z}_s) - F_s(\bar{z}_s)|^2) - \gamma|\tilde{Y}_s|^2\right) ds\right]. \end{aligned}$$

Fix $\gamma > (1+\theta)\theta$ and $q = (1+\theta)\theta/\gamma$. Let $\delta = \gamma/(1+\theta)$. Then

$$\begin{aligned} \mathbb{E}\left[e^{\gamma t}|\tilde{Y}_t|^2 + \int_t^T e^{\gamma s}|\tilde{Z}_s|^2\, ds\right] &\leq \mathbb{E}\left[\int_t^T e^{\gamma s} q\left(|\tilde{z}_s|^2 + \frac{1}{\theta}|\bar{F}_s(\bar{z}_s) - F_s(\bar{z}_s)|^2\right) ds\right] \\ &\leq q\|\tilde{z}\|^2_{\mathbb{H}^2_\gamma} + \frac{q}{\theta}\|\bar{F}(\bar{z}) - F(\bar{z})\|^2_{\mathbb{H}^2_\gamma}. \end{aligned}$$

This concludes the proof of the lemma. □

Lemma A.8. *Let $\bar{F} : \Omega \times [0,T] \times \mathbb{R}^d \to \mathbb{R}$ be a measurable function and let F satisfies the hypotheses of Lemmas A.4 and A.5. Fix $\xi \in L^2(\Omega, \mathcal{F}_T)$. Let $\bar{z}, z, \bar{Z}, Z \in \mathbb{H}^2$ and $\bar{Y}, Y \in \mathcal{S}^2$ be such that*

$$\bar{Y}_t = \xi + \int_t^T \bar{F}_s(\bar{z}_s, \bar{Z}_s)\, ds - \int_t^T \bar{Z}_s\, dW_s\,, \;\; t \in [0,T]\,,$$

and

$$Y_t = \xi + \int_t^T F_s(z_s, Z_s)\, ds - \int_t^T Z_s\, dW_s\,, \;\; t \in [0,T]\,.$$

Then there is $\gamma > 0$ and $q \in (0,1)$ such that for $t \in [0,T]$ we have

$$e^{\gamma t}\mathbb{E}|\bar{Y}_t - Y_t|^2 + \|\bar{Z} - Z\|^2_{\mathbb{H}^2_\gamma} \le q\|\bar{z} - z\|^2_{\mathbb{H}^2_\gamma} + \frac{q}{\max(K,\theta)}\|\bar{F}_s(\bar{z}_s, \bar{Z}_s) - F_s(\bar{z}_s, \bar{Z}_s)\|^2_{\mathbb{H}^2_\gamma}.$$

Proof. Consider $\gamma > 0$ which we will fix later. We denote $\tilde{Y} := \bar{Y} - Y$, $\tilde{Z} := \bar{Z} - Z$, and $\tilde{z} := \bar{z} - z$. We then apply Itô's formula to $e^{\gamma t}|\tilde{Y}_t|^2$:

$$\begin{aligned} e^{\gamma t}|\tilde{Y}_t|^2 &+ \int_t^T e^{\gamma s}|\tilde{Z}_s|^2\, ds \\ &= \int_t^T e^{\gamma s}(2\tilde{Y}_s\,(\bar{F}_s(\bar{z}_s, \bar{Z}_s) - F_s(z_s, Z_s)) - \gamma|\tilde{Y}_s|^2)\, ds \\ &\quad - 2\int_t^T e^{\gamma s}\tilde{Z}_s\,\tilde{Y}_s\, dW_s\,. \end{aligned}$$

Due to Remark A.3, the stochastic integral vanishes by taking expectation. Hence

$$\begin{aligned} &\mathbb{E}\left[e^{\gamma t}|\tilde{Y}_t|^2 + \int_t^T e^{\gamma s}|\tilde{Z}_s|^2\, ds\right] \\ &\quad = \mathbb{E}\left[\int_t^T e^{\gamma s}(2\tilde{Y}_s\,(\bar{F}_s(\bar{z}_s, \bar{Z}_s) - F_s(z_s, Z_s)) - \gamma|\tilde{Y}_s|^2)\, ds\right]. \end{aligned}$$

Notice that by assumptions of the lemma for all $s \in [t,T]$ it holds that

$$\begin{aligned} |\bar{F}_s(\bar{z}_s, \bar{Z}_s) - F_s(z_s, Z_s)| &\le |\bar{F}_s(\bar{z}_s, \bar{Z}_s) - F_s(\bar{z}_s, \bar{Z}_s)| + |F_s(\bar{z}_s, \bar{Z}_s) - F_s(z_s, Z_s)| \\ &\le |\bar{F}_s(\bar{z}_s, \bar{Z}_s) - F_s(\bar{z}_s, \bar{Z}_s)| + \theta|\bar{z}_s - z_s| + K|\bar{Z}_s - Z_s|\,. \end{aligned}$$

Then by the Young inequality for any $\delta > 0$, we have

$$\begin{aligned} &\mathbb{E}\left[e^{\gamma t}|\tilde{Y}_t|^2 + \int_t^T e^{\gamma s}|\tilde{Z}_s|^2\, ds\right] \\ &\le \mathbb{E}\left[\int_t^T e^{\gamma s}(2|\tilde{Y}_s|(|\bar{F}_s(\bar{z}_s, \bar{Z}_s) - F_s(\bar{z}_s, \bar{Z}_s)| + \theta|\bar{z}_s - z_s| + K|\bar{Z}_s - Z_s|) - \gamma|\tilde{Y}_s|^2)ds\right] \\ &\le \mathbb{E}\Bigg[\int_t^T e^{\gamma s}\Big((\theta + K + 1)\delta|\tilde{Y}_s|^2 + \delta^{-1}(K|\tilde{Z}|^2 + \theta|\tilde{z}_s|^2 + |\bar{F}_s(\bar{z}_s, \bar{Z}_s) - F_s(\bar{z}_s, \bar{Z}_s)|^2) \\ &\qquad - \gamma|\tilde{Y}_s|^2\Big)ds\Bigg]. \end{aligned}$$

Let us take $\gamma > 0$ sufficiently large so that $\tilde{q} := \max(\frac{(1+\theta+K)K}{\gamma}, \frac{(1+\theta+K)\theta}{\gamma}) \in (0, 1/2)$. Let $\delta := \gamma/(1+\theta+K)$ so that

$$\begin{aligned}
&\mathbb{E}\left[e^{\gamma t}|\tilde{Y}_t|^2 + (1-\tilde{q})\int_t^T e^{\gamma s}|\tilde{Z}_s|^2\,ds\right] \\
&\quad\le \mathbb{E}\left[\int_t^T e^{\gamma s}\tilde{q}\left(|\tilde{z}_s|^2 + \frac{1}{\max(K,\theta)}|\bar{F}_s(\bar{z}_s,\bar{Z}_s) - F_s(\bar{z}_s,\bar{Z}_s)|^2\right)ds\right] \\
&\quad\le \tilde{q}\|\tilde{z}\|^2_{\mathbb{H}^2_\gamma} + \frac{\tilde{q}}{\max(K,\theta)}\|\bar{F}_s(\bar{z}_s,\bar{Z}_s) - F_s(\bar{z}_s,\bar{Z}_s)\|^2_{\mathbb{H}^2_\gamma}.
\end{aligned}$$

Dividing by $(1-\tilde{q}) \in (1/2, 1)$ we obtain

$$\mathbb{E}\left[e^{\gamma t}|\tilde{Y}_t|^2 + \int_t^T e^{\gamma s}|\tilde{Z}_s|^2\,ds\right] \le q\|\tilde{z}\|^2_{\mathbb{H}^2_\gamma} + \frac{q}{\max(K,\theta)}\|\bar{F}_s(\bar{z}_s,\bar{Z}_s) - F_s(\bar{z}_s,\bar{Z}_s)\|^2_{\mathbb{H}^2_\gamma},$$

where $q := \frac{\tilde{q}}{1-\tilde{q}}$. Since $0 < \tilde{q} < 1/2$ we have that $q \in (0,1)$. □

A.1. **BSDE with drivers of quadratic growth.** Since we are using BSDE theory in the proof of the main result, we would like to present some results on BSDE with drivers of quadratic growth. We refer to [16].

Consider the following system:

$$\begin{aligned}
X_t &= x + \int_0^t b(s, X_s)\,ds + \int_0^t \sigma(s, X_s)\,dW_s\,, \\
Y_t &= g(X_T) + \int_t^T f(s, X_s, Z_s)\,ds - \int_t^T Z_s\,dW_s.
\end{aligned} \tag{49}$$

Theorem A.9 (Theorem 3.6 in [16]). *Let $b : [0,T]\times\mathbb{R}^d \to \mathbb{R}^d$ and $\sigma : [0,T]\times\mathbb{R}^d \to \mathbb{R}^{d\times d'}$ be Lipschitz continuous with Lipschitz constant C and $|b(t,0)| \le C$ and $|\sigma(t,0)| \le C$ for all $t \in [0,T]$. Let $g : \mathbb{R}^d \to \mathbb{R}$ and $f : [0,T]\times\mathbb{R}^d\times\mathbb{R}^d \to \mathbb{R}$ be measurable functions and let us assume that there exists constant C such that for all $r \in \mathbb{R}^+, t \in [0,T], x, x' \in \mathbb{R}^d$, and $z, z' \in \mathbb{R}^d$*

$$\begin{aligned}
&|f(t,x,z)| \le C(1+|z|^2)\,, \\
&|g(x)| \le C\,, \\
&|f(t,x,z) - f(t,x,z')| \le C(1+|z|+|z'|)|z-z'|\,, \\
&|f(t,x,z) - f(t,x',z)| \le C(1+|x|^r+|x'|^r)|x-x'|\,, \\
&|g(x) - g(x')| \le C(1+|x|^r+|x'|^r)|x-x'|\,.
\end{aligned}$$

There exists a solution (Y,Z) of the Markovian BSDE (49) in $\mathcal{S}^2\times\mathbb{H}^2$ and this solution is unique among solutions $(Y,Z) \in \mathcal{S}^2\times\mathbb{H}^2$ such that Y is bounded. Moreover, we have

$$|Z_t| \le C(1+|X_t|^{1+r})\ \forall t \in [0,T]\,,$$

and

$$\|Z \bullet W\|_{BMO} < \infty\,,$$

where

$$\|M\|_{BMO} := \sup_{\tau\in\mathcal{T}_0^T} \|\mathbb{E}[\langle M\rangle_T - \langle M\rangle_\tau \mid \mathcal{F}_\tau]\|_\infty\,;$$

here the supremum is taken over all stopping times in $[0,T]$.

References

[1] R. Bellman, Functional equations in the theory of dynamic programming. V. Positivity and quasi-linearity, *Proc. Natl. Acad. Sci. USA.*, 41 (1955), pp. 743-746.

[2] R. Bellman, *Dynamic Programming*, Princeton University Press, Princeton, NJ, USA, 1957.

[3] R. A. Howard, *Dynamic Programming and Markov Processes*, MIT Press, Cambridge, MA, 1960.

[4] M.L. Puterman and S. L. Brumelle, On the convergence of policy iteration in stationary dynamic programming, *Math. Oper. Res.*, 4 (1979), pp. 60-69.

[5] M. L. Puterman, On the convergence of policy iteration for controlled diffusions, *J. Optim. Theory Appl.*, 33 (1981), pp. 137–144 .

[6] N. V. Krylov, *Controlled Diffusion Processes*, Springer, New York, 1980.

[7] O. Hernandez-Lerma and J. Lasserre, *Discrete-Time Markov Control Processes*, Springer, New York, 1996.

[8] R. Buckdahn and S. Peng, Ergodic Backward SDE and Associated PDE, R.C. Dalang, M. Dozzi and F. Russo eds., *Progr. Probab. 45*, Birkhäuser, Basel, 1999.

[9] M. S. Santos and J. Rust, Convergence properties of policy iteration. *SIAM J. Control and Optim.*, 42 (2004), pp 2094–2115, .

[10] M. Fuhrman and G. Tessitore, Infinite horizon backward stochastic differential equations and elliptic equations in Hilbert Spaces, *Ann. Probab.*, 32 (2004), pp. 607–660.

[11] W. H. Fleming and H. M. Soner, *Controlled Markov Processes and Viscosity Solutions*, Stoch. Model. Appl. Probab. 25, Springer, New York, 2006.

[12] H. Dong and N. V. Krylov, The rate of convergence of finite-difference approximations for parabolic Bellman equations with Lipschitz coefficients in cylindrical domains, *Appl. Math. Optim.* 56 (2007), pp. 37–66.

[13] O. Bokanowski, S. Maroso, and H. Zidani, Some convergence results for Howard's algorithm, *SIAM J. on Numer. Anal.*, 47 (2009), pp. 3001-3026.

[14] I. Gyöngy and D. Šiska, On finite-difference approximations for normalized Bellman equations, *Appl. Math. Optim.*, 60 (2009), pp. 297–339.

[15] H. Pham. *Continuous-time Stochastic Control and Optimization with Financial Applications*, Springer, New York, 2009.

[16] A. Richou, Markovian quadratic and superquadratic BSDEs with an unbounded terminal condition, *Stochastic Process. Appl.*, 122 (2012), pp. 3173-3208.

[17] N. Bäuerle and U. Rieder, Control improvement for jump-diffusion processes with applications to finance, *Appl. Math. and Optim.*, 65 (2012), pp. 1-14.

[18] S. D. Jacka and A. Mijatović, On the policy improvement algorithm in continuous time, *Stochastics*, 89 (2017), pp. 348–359.

[19] S. D. Jacka, A. Mijatović and D. Siraj, Coupling and a Generalised Policy Iteration Algorithm in Continuous Time, *arXiv*:1707.07834, 2017.

[20] J. Maeda and S. D. Jacka, Evaluation of the Rate of Convergence in the PIA. *arXiv*:1709.06466, 2017.